Actes des 9èmes Journées Internationales de la Linguistique de corpus

3-6 juillet 2017

Grenoble, France

LIDILEM Université Grenoble Alpes

UMR 5316 LITT & ARTS

ILCEA4 Université Grenoble Alpes

LIG

Comités

Comité scientifique

ANTOINE Jean-Yves (Université de Tours, LI)
ARAUJO Silvia (University of Minho, Portugal)
BONDI Marina (Università degli Studi di Modena e Reggio Emilia, Italie)
BOULTON Alex (Université de Lorraine, ATILF)
BOUTET Dominique (Université de Rouen, DySoLa)
CARTER-THOMAS Shirley (Institut Mines- Télécom, LaTTiCe)
DE GIOVANNI Cosimo (Université de Cagliari, Italie)
DIWERSY Sascha (Université Paul Valéry, Praxiling)
DOSTIE Gaétane (Université de Sherbrooke, CANADA)
ESKHOL Iris (Université d'Orléans, LLL)
ETIENNE Carole (ICAR, CNRS)
FONTENELLE Thierry (Translation Centre for the Corpora of the European Union)
GASIGLIA Nathalie (Université de Lille3, STL)
GRANGER Sylviane (Université Catholique de Louvain, Belgique)
HUNSTON Susan (University of Birmingham)
JACQUES Marie-Paule (Université Grenoble-Alpes, Lidilem)
KRAIF Olivier (Université Grenoble-Alpes, Lidilem)
KUBLER Natalie (Université Paris Diderot, CLILLAC)
LANDRAGIN Frédéric (LaTTiCe, CNRS)
LEBARBE Thomas (Université Grenoble-Alpes, Litt&Arts)
MANIEZ François (Université Lyon II, CRTT)
MAUREL Denis (Université de Tours, LI)
NESI Hilary (Coventry University, Grande-Bretagne)
NEVEU Franck (Université Paris-Sorbonne Paris IV, ILF)
PARISSE Christophe (MoDyCo, Inserm)
PETRAUSKAITE Rūta (Vytautas Magnus University, Lituanie)
PIERREL Jean-Marie (Université de Lorraine, ATILF)
POUDAT Céline (Université de Nice, BCL)
RASTIER François (UMR 7114 CNRS)
SCHNEDECKER Catherine (Université de Strasbourg, LiLPa)
SIEPMANN Dirk (Université d'Osnabrück, Allemagne)
SIMON Anne-Catherine (Université catholique de Louvain, Belgique)
TEUBERT Wolfgang (University of Birmingham)
TODIRASCU Amalia (Université de Strasbourg, LiLPa)
TUTIN Agnès (Université Grenoble-Alpes, LIDILEM)
WILLIAMS Geoffrey (Université de Bretagne-Sud, Litt&Arts)

Comité d'organisation

Tatiana Aleksandrova (Université de Grenoble-Alpes-Alpes, LIDILEM)
Georges Antoniadis (UUniversité de Grenoble-Alpes-Alpes, LIDILEM)
Emmanuelle Esperança-Rodier (UUniversité de Grenoble-Alpes-Alpes, LIG)
Cécile Frérot (Université de Grenoble-Alpes, ILCEA4)
Laure Gardelle (Université de Grenoble-Alpes, LIDILEM)
Francis Grossmann (Université de Grenoble-Alpes, LIDILEM)
Laura Hartwell (Université de Grenoble-Alpes, LIDILEM)
Marie-Paule Jacques (Université de Grenoble-Alpes, LIDILEM)
Olivier Kraif (Université de Grenoble-Alpes, LIDILEM)

Thomas Lebarbé (Université de Grenoble-Alpes, Litt&Arts)
Samia Ounoughi (Université de Grenoble-Alpes, LIDILEM)
Claude Ponton (Université de Grenoble-Alpes, LIDILEM)
Solange Rossato (Université de Grenoble-Alpes, LIG)
Isabelle Rousset (Université de Grenoble-Alpes, LIDILEM)
Julie Sorba (Université de Grenoble-Alpes, LIDILEM)
Agnès Tutin (Université de Grenoble-Alpes, LIDILEM)
Geoffrey Williams (Université de Grenoble-Alpes, Litt&Arts)
Virginie Zampa (Université de Grenoble-Alpes, LIDILEM)

Partenaires

Table des matières

Plénières

Corpus multimodaux de Langues Vocales et de Langues des Signes enregistrés en capture de mouvement, vers un changement paradigmatique

Dominique Boutet
Université de Rouen, France

L'analyse de la multimodalité repose sur une conception spécifique de la gestuelle. Celle-ci, grandement inspirée par l'imagerie (McNeill 1992, 2000, Cuxac 2000, Taub 2001), s'ancre pour l'essentiel sur une tradition cognitive (Fauconnier 1984). Cette approche imagique du sens dans la gestualité symbolique trouve des racines philosophiques, mais ne fait jamais référence à l'histoire de l'art, alors qu'une réflexion extrêmement riche et diverse la parcourt. Les artistes et leurs commentateurs apportent non seulement des évidences sur l'absence de simplification représentationnelle de l'image peinte ou sculptée à son référent supposée, mais, de plus, révèle l'ensemble des propositions faites pour dépasser l'immobilité, le figement des gestes en autant de postures qu'imposent la toile ou le bloc de marbre. Nous montrerons que cette conception figée de la gestualité est toujours à l'œuvre dans les analyses de la gestualité coverbale et des langues signées, rendant statique ce qui est en mouvement, oublieuse du corps et de sa matérialité au profit de sa représentation à la surface d'un écran. Les techniques de capture du mouvement sont en train de réaliser une rupture technologique dans l'enregistrement des gestes. Elles captent le mouvement dans l'espace selon ses trois dimensions, et plus seulement l'image projetée en deux dimensions. On a désormais accès aux données mêmes des gestes et plus seulement à leurs images. Ces technologies nous mettent en demeure de réviser totalement notre conception de la gestualité, en la débarrassant des oripeaux de son figement en postures, en sortant d'une pure conception imagique, en étendant le cadre de référence égocentré à des cadres intrinsèques. Un changement paradigmatique est en cours. Si les études gestuelles sont restées fixées sur l'image statique à l'heure de leur animation (l'apparition de la vidéo grand public), il est urgent, désormais, qu'elles rattrapent ces nouveaux modes d'enregistrement et qu'elles se hissent à ce qu'elles mettent au jour, c'est-à-dire à la compréhension des mouvements conjoints des différentes parties du corps ; en somme au dépassement du mouvement pour comprendre comment le corps produit son expressivité mobile.

Collocation and opportunities for interdisciplinarity

Michaela Mahlberg
University of Birmingham, Royaume-Uni

The concept of 'collocation' is one of the most fundamental concepts in corpus linguistics. It is often considered in relation to work by Firth (1957) who highlights that habitual co-occurrence patterns are crucial to the meaning of a word. Sinclair et al. (2004: 10) define 'collocation' as "the co-occurrence of two items in a text within a specified environment". Corpus software packages tend to include the retrieval of collocates among their standard functionalities. Although the concept seems to have been around for a long time, collocation still has the potential to move corpus linguistics forward by raising new questions and furthering research across disciplinary boundaries. To make this point, I will consider examples from the CLiC project1 and the analysis of literary texts. A new functionality of the CLiC app (http://clic.bham.ac.uk/ Mahlberg et al. 2016, Stockwell and Mahlberg 2015) is a KWICgrouping option (cf. O'Donnell 2008). I will demonstrate how the KWICgrouper can support the viewing of collocations in context. I will also discuss challenges of comparing collocations across corpora and propose some methodological solutions to this problem. Against this background, I will indicate potential for future work in the area of collocation studies.

Références bibliographiques :

Firth, J.R.,2011, *Papers in Linguistics 1934-51*London Oxford University Press.

Mahlberg, M., Stockwell, P., de Joode, J., Smith, C., O'Donnell, M. Brook, 2016, CLiC Dickens – Novel uses of concordances for the integration of corpus stylistics and cognitive poetics, *Corpora*, 11 (3), 433-463. http://www.euppublishing.com/doi/pdfplus/10.3366/cor.2016.0102

O'Donnell, M. B. 2008, KWICgrouper – designing a tool for corpus-driven concordance analysis, *International Journal of English Studies*, 8 (1), 107-121. http://revistas.um.es/ijes/article/view/49121/46991

Sinclair, J., Jones, S. & Daley, R. , 2004,*English Collocation Studies. The OSTI Report.*,(ed. by R. Krishnamurthy). London: Bloomsbury.

Stockwell, P & Mahlberg, M., 2015, Mind-modelling with corpus stylistics in *David Copperfield*, *Language and Literature*, 24 (2), 129-147. http://journals.sagepub.com/doi/pdf/10.1177/0963947015576168

Présentations orales

Session 1.A.
Constitution, annotation de corpus

Corpus oraux et sociolinguistique des interactions au travail : de la constitution à l'analyse

Anouchka Divoux
ATILF UMR 7118 – Université de Lorraine
CNRS ATILF – 44 avenue de la libération, 54000 NANCY
anouchka.divoux@univ-lorraine.fr

1. Introduction

Dans le cadre de notre étude visant à saisir les caractéristiques (socio)linguistiques et interactionnelles des questions au sein des réunions de travail en entreprise, nous avons été amenée à constituer un corpus de données orales au sein d'une entreprise du domaine de la construction et de la rénovation de patrimoine. En d'autres termes, notre étude cherche à analyser les rapports entre l'utilisation des questions et les différents éléments extralinguistiques du contexte, notamment le statut des locuteurs, leur rôle interactionnel, leur sexe, mais aussi le contexte économique et social de l'entreprise. Pour constituer notre corpus de travail, nous avons fait le choix de réunir des données écologiques (Bert et al, 2010), autrement dit, des données recueillies « en situation naturelle, c'est-à-dire dans le contexte social ordinaire des activités documentées, en dehors de toute élicitation ou expérimentation » (Bert et al, 2010, p.17). Néanmoins, en dehors des phases de collecte et d'analyse, d'autres étapes jouent un rôle crucial à la fois parce qu'elles peuvent éclairer les données linguistiques recueillies, mais aussi parce qu'elles permettent de questionner notre posture de chercheur. Ainsi, notre communication portera sur les aspects méthodologiques de la constitution d'un corpus de données orales en situation de travail (de la recherche du terrain à l'analyse des données) ainsi que sur le statut des données recueillies.

2. De la recherche d'un terrain...

Dans cette première partie, nous évoquerons à la suite de André (2006), les difficultés rencontrées pour nous faire accepter en tant que chercheuse en Sciences du Langage dans une entreprise. En effet, les enjeux économiques et financiers sont souvent tels qu'il est impossible de prétendre enregistrer des données dans une entreprise sans éveiller des soupçons d'espionnage industriel ou d'immixtion dans la complexité des relations sociales internes de l'entreprise. La seconde difficulté est relative à notre statut de linguiste. De ce fait, les salariés craignent parfois un jugement négatif sur leurs productions langagières. En conséquence, sans commande de recherche de la part de l'entreprise et sans allié (au sens de Beaud et Weber, 2010) au sein de celle-ci, il est presque impossible de s'y frayer un chemin.

3... A l'exploration préliminaire

La deuxième étape que nous évoquerons portera sur la prise de contact avec l'entreprise, une fois le terrain de recherche trouvé. Celle-ci s'avère indispensable sur deux plans. D'une part, elle permet la familiarisation avec le terrain (Garfinkel, 1967). Dans la mesure où nous effectuons une recherche sociolinguistique, cette exploration préliminaire du terrain est cruciale car elle permet de comprendre le fonctionnement général de l'entreprise, mais aussi de clarifier les fonctions, le statut et les relations interpersonnelles de ses différents acteurs. Cette étape nous paraît d'autant plus incontournable dans le cadre de notre étude. Celle-ci prenant place dans une entreprise de restauration de patrimoine et de construction, un technolecte précis est lié à l'activité de travail qui n'est que difficilement accessible aux personnes non initiées. D'autre part, cette période de prise de contact permet l'instauration d'un climat de confiance auprès des salariés qui consentent alors plus facilement à échanger et à répondre à nos nombreuses interrogations. Sans cette période d'immersion préalable, la majorité du corpus resterait inacessible tant en termes de lexique utilisé (sigles, technolecte) que de compréhension du contexte.

4. Aspects éthiques de la constitution du corpus

Notre implication en tant que chercheur sur un terrain comme celui de l'entreprise oblige nécessairement à une réflexion préliminaire sur les conditions d'enquête et surtout sur les conditions de notre présence. Dès l'instant où une entreprise ouvre ses portes à un chercheur, se pose la question de la responsabilité morale vis-à-vis de celle-ci. Ainsi, un certain nombre de garanties et de précautions semblent incontournables : garantie de confidentialité, anonymisation de certaines données (patronymes, noms de l'entreprise et des partenaires) mais aussi partage des objectifs et des résultats de la recherche (Spradley 1979, Feldman, 2000). Outre l'aspect éthique de la recherche, ce type d'engagements semble rassurer les interactants. En conséquence, de telles précautions nous permettent un recueil de productions orales moins surveillées et de fait, plus authentiques.

5. Recueil et traitement des données

De même, l'étape de recueil amène à une réflexion sur la posture du chercheur. Il nous semblait impératif de pouvoir être présente lors de l'enregistrement des données pour deux raisons. D'une part pour des raisons pratiques : le nombre de participants étant élevé (jusqu'à une dizaine), l'attribution des bons tours de paroles aux différents locuteurs aurait pu être compromise lors de la phase de transcription. D'autre part, si le travail de transcription est un processus long et laborieux (compter environ une heure de transcription pour une minute de polylogue), notre présence lors des réunions, ainsi que les prises de note effectuées (premiers et derniers mots de chaque locuteur, événements extralinguistiques, interrogations en suspens) ont permis de faciliter la compréhension et l'interprétation de certains segments durant la phase de transcription, mais aussi d'analyse. Cependant, nous avons tâché d'adopter une posture qui soit la plus discrète possible (hors du champ de vision et hors du cercle des participants), à l'image de la participation passive décrite par Duranti (1997). Cette position nous a semblé minimiser les effets du paradoxe de l'observateur (Labov, 1972). Nous sommes consciente que les effets de la présence de l'observateur sur les participants ne peuvent être totalement annulés. Néanmoins, il est important de tenir compte de ceux-ci s'ils sont décelables. Toutefois, malgré notre présence, nous avons pu constater que « les participants ont des objectifs précis à atteindre pendant les réunions, par conséquent, les exigences de fonctionnement du groupe semblent primer sur les éventuels effets de l'observation » (André, 2006).

6. Un retour sur le terrain nécessaire

Si la première phase d'exploration est nécessaire, un retour sur le terrain est indispensable après la phase de constitution en vue de faciliter l'interprétation de certains segments (mais aussi à fortiori, l'analyse). Outre la phase de recueil stricte, nous avons passé quantité de moments informels auprès des participants afin de compléter nos connaissances et de mieux comprendre les éléments en jeu lors des réunions. Ceux-ci ont souvent permis de jeter un éclairage nouveau sur certains de leurs propos qui nous étaient obscurs en raison des nombreuses connaissances partagées (Gumperz, Hymes, 1972) auxquelles nous n'avions pas accès (chantiers en cours ou passés, anecdotes, sigles et acronymes). Au cours d'une précédente recherche (Divoux, 2016), nous avons été amenée à analyser certaines des questions produites en réunions comme des actes de langage pragmatiquement hybrides (Kerbrat, 2008). En effet, outre la stricte demande d'information généralement associée aux questions, celle-ci peuvent, lors des réunions de travail, poursuivre un second but pragmatique. Celui-ci est intimement lié aux paramètres du contexte comme le statut des locuteurs, leur rôle interactionnel, l'activité de travail mais aussi le but de la réunion :

- volonté de faciliter l'intercompréhension : « est-ce que c'est bon pour tout le monde ? ».
- mise en difficulté d'un interlocuteur : un supérieur hiérarchique demandant des comptes à un membre de son équipe : « alors je me demande pourquoi tu y étais ».
- expression de l'expertise : voir infra.

C'est à ce type de questions hybrides que nous nous intéressons ici. Dans l'exemple suivant, les participants d'un service effectuent un tour de table pour décrire leurs activités de la semaine. P6 évoque ses activités actuelles et évoque des problèmes liés à un type de transformateur électrique (TCT). A la suite de son intervention, P5 pose une question pour préciser le type de transformateur (EJ33).

P6	et sinon en en termes de prévention on est sur un autre sujet + que je que je pilote aussi c'est sur euh vous savez les les les transfo de : : // les les les T- enfin ce qu'on appelle les TCT alors je sais pas si ça vous parle les mes- les mesureurs de tension euh en 400 et 220 000 on a des des problèmes de dérives sur ces appareils et après euh ça se finit pas bien pour eux [. . .] c'est euh la partie maintenance euh où on veut améliorer ce qu'on appelle les plans de préventions qui permettent euh de dé- définir les mesures quand on fait /le ;les/ chantiers
P5	**c'est les fameux EJ33 ↗ non ça n'a rien à voir ↘**
P6	alors c'est pas les EJ33 <parce que les EJ33 c'est sur le
P5	c'est une autre phase ça> ↗
P6	63 000 volts
P5	ouais

Extrait du corpus réunion 1

La question posée par P5 (en gras) aurait pu être considérée comme une simple demande d'information pour préciser le type de transformateur électrique dont il s'agissait. Toutefois, plusieurs éléments nous ont permis de l'analyser comme une démonstration d'expertise (Ford, 2008) de la part de P5. En effet, cette réunion se déroule au sein du service des ressources humaines de l'entreprise et les salariés du service (mis à part P6) n'ont généralement pas connaissance de ce type de technolecte (EJ33), utilisé majoritairement par les membres du service maintenance. Seulement, P5 est présente dans l'entreprise depuis une vingtaine d'années et a été amenée à côtoyer différents services, ce qui lui a permis d'acquérir des nombreuses connaissances, notamment en termes de technolectes. En posant la question « c'est les fameux EJ33 ↗ non ça n'a rien à voir ↘ », elle demande d'une part à P6 une confirmation ou une infirmation et d'autre part, elle fait valoir une certaine maîtrise du technolecte d'un autre service. Cette interprétation a été rendue possible grâce aux nombreux échanges menés a posteriori avec les locuteurs présents lors de cette réunion.

7. Conclusion

La constitution d'un corpus oral en entreprise dans le cadre d'une étude sociolinguistique suppose une méthodologie de recueil mûrement réfléchie en amont. Ce type de réflexion sur les pratiques de recueil semble incontournable tant d'un point de vue éthique que pratique car elle permet de désamorcer les éventuelles inquiétudes des salariés et donc, garantit un recueil de données plus authentiques. Par ailleurs, nos réflexions soulignent l'importance du travail du chercheur dans le recueil. Si les données lui sont accessibles et compréhensibles, c'est grâce au travail d'immersion en plusieurs temps qu'il a lui-même mené, parfois sur de longues périodes. Enfin, si les efforts fournis par le chercheur pour se faire accepter dans un milieu généralement réfractaire à sa présence sont conséquents, ceux-ci s'avèrent importants pour gagner la confiance des enquêtés et réunir des informations précieuses pour l'analyse des données.

Références bibliographiques

Bert, M., Bruxelles, S., Etienne, C., Jouin-Chardon, E., Lascar, J., Mondada, L., ... Traverso, V. (2010). Grands corpus et linguistique outillée pour l'étude du français en interaction (plateforme CLAPI et corpus CIEL), *Pratiques. Linguistique, littérature, didactique,*147/148, 17-34.

André, V. (2006). *Construction collaborative du discours au sein de réunions de travail en entreprise : de l'analyse micro-linguistique à l'analyse socio-interactionnelle*, (Thèse de doctorat). Université de Nancy 2.

Beaud, S. & Weber, F. (2010). *Guide de l'enquête de terrain : produire et analyser des données ethnographiques.*, Paris : La Découverte.

Divoux, A. (2016). *Analyse des aspects linguistiques, praxéologiques et genrés de la question en réunion de travail.* (Mémoire de Master 2). Université de Lorraine, Nancy.

Duranti, A. (1997). *Linguistic Anthropology.* Cambridge : Cambridge University Press.

Feldman, J. (2000). *L'éthique dans la pratique des sciences humaines : dilemmes.* Paris : L'Harmattan.

Ford, C. E. (2008). *Women speaking up : Getting and using turns in workplace meetings.* Londres : Palgrave Macmillan.

Garfinkel, H. (1967). *Studies in ethnomethodology.* Upper Saddle River : Prentice-Hall.

Gumperz, J., Hymes, D. (1972). *Directions in sociolinguistics. The ethnography of communication.* Oxford : Basil Blackwell.

Kerbrat-Orecchioni, C. (2008). *Les actes de langage dans le discours : théorie et fonctionnement (2ème édition).* Paris : Armand Colin.

Labov, W. (1972). Some principles of linguistic methodology. *Language in society*, 1(01), 97-120.

Spradley, J. P. (1979). *The ethnographic interview.* NewYork : Holt, Rinehart & Winston.

Le regard et les mains : Annotation et analyse multipistes d'un corpus de LSF

Annelies Braffort
LIMSI CNRS, Université Paris-Saclay
annelies.braffort@limsi.fr

1. Introduction

Les LS sont des langues naturelles pratiquées au sein des communautés de Sourds et la Langue des Signes Française (LSF) est celle utilisée en France. Ce sont des langues visuo-gestuelles : une personne s'exprime en LS en utilisant de nombreuses composantes corporelles (les mains et les bras, mais aussi les expressions du visage, le regard, le buste, etc.) et son interlocuteur perçoit le message par le canal visuel. Le système linguistique des LS exploite ces canaux spécifiques : de nombreuses informations sont exprimées simultanément et s'organisent dans l'espace, et l'iconicité joue un rôle central. A ce jour, les LS n'ont pas de système d'écriture ni de système graphique standard pour la transcription. Elles sont encore peu décrites et peu dotées et la constitution de corpus répond à un besoin crucial, pour les études en linguistique mais aussi en traitement automatique. L'annotation et l'analyse de corpus de LSF nécessite la manipulation de partitions d'annotation, ou chaque composante est annotée sur une piste différente. Lors de l'analyse, il faut pouvoir mettre en relation des segments d'annotation distribués sur plusieurs pistes et plus ou moins synchronisés. Cette contribution décrit une méthode d'analyse permettant de catégoriser les relations temporelles entre segments issus de deux pistes d'annotation différentes. La méthode a été utilisée pour étudier les relations temporelles entre les activités du regard et des mains lors de la production d'unités gestuelles. Après avoir précisé les phénomènes étudiés et mis en relation, je présenterai la méthodologie mise en place pour étudier le phénomène comme un tout, en commençant par décrire le corpus, les annotations réalisés, la méthode d'analyse, puis des résultats.

2. Le rôle du regard en LSF

Dans les LS, l'activité non manuelle peut être combinée avec d'autres composantes, manuels ou non manuels, et assurer des rôles grammaticaux variés. C'est typiquement le cas pour le regard (Engberg-Pedersen, 1999) : certains signes lexicaux peuvent être accompagnés par une direction de regard spécifique, certaines unités gestuelles nécessitent un regard dirigé vers l'espace de signation (par exemple le pointage). Dans certains modèles théoriques (Cuxac 2000), le regard a une fonction sémiotique, permettant de distinguer deux modes d'expression : sans ou avec visée illustrative (regard dirigé vers l'interlocuteur ou non). Mais c'est sans doute aussi pour des raisons de contraintes physiologique que des directions de regard spécifiques sont parfois associées à des mouvements des mains. Parmi les unités gestuelles susceptibles d'être accompagnées par une activité particulière du regard, on trouve en particulier des signes non lexicaux dont la fonction est de décrire la taille et la forme d'entités (lieux, objets, personnes) dans un discours. Ces signes sont nommées, selon les théories linguistiques, « Transfert de Taille et de Forme » (Cuxac, 2000), ou encore « Size and Shape Depicting Signs » (Johnston 2014). Ces descriptions ne comportent ni procès, ni actant. D'après Cuxac, la structure de ces signes, que l'on nommera TTF dans la suite, est la suivante : les mains se mettent en place dans l'espace, puis le regard fixe un point de l'espace et précède de peu le déploiement du signe. La question est de savoir si ces phénomènes sont systématiques et d'en mesurer les valeurs (durée, direction). On va se concentrer ici sur la propriété de précédence du regard sur le déploiement du signe.

3. Constitution et annotation du corpus

Dans le but de disposer de données permettant des études dans plusieurs disciplines (sciences du mouvement, du langage, traitement automatique), nous avons constitué un corpus de LSF, nommé MOCAP1 (Benchiheub, Berret, Braffort 2016). Ce corpus a été capté à l'aide d'un système de capture de mouvement (10 caméras OptiTrack et 40 marqueurs sur le corps et le visage) et une caméra HD (figure 1). Nous disposons ainsi de données vidéo nous permettant d'annoter le corpus avec des logiciels

d'annotation classiques, mais aussi de données 3d afin de pouvoir effectuer des mesures précises pour décrire quantitativement les phénomènes étudiés. Ce corpus comporte divers types de productions (description, narration, explication, brèves journalistiques), obtenus à l'aide de matériaux d'élicitation variés, structurées en cinq taches différentes. Huit locuteurs ont participé à ce corpus, quatre hommes et quatre femmes, de profils socio-linguistiques divers. La première tache a été annotée. Cette tache consistait pour chaque locuteur à décrire 25 photos qui lui étaient montrées l'une après l'autre (figure 2). Nous disposons donc de 200 descriptions, qui ont été annotées à l'aide du logiciel d'annotation ANVIL, qui permet l'affichage en 3D de données de mocap (Héloir et al, 2010). Le schéma d'annotation comporte une piste pour le regard, une pour chaque main, et une pour l'annotation globale des unité gestuelles (figure 3). Les pistes Regard et Mains comportent des annotation sur la forme et la piste relative aux Signes, sur la fonction. Regard : nous avons rapporté dans Braffort (2014) certaines des pratiques pour l'annotation de l'activité du regard et montré que le niveau d'abstraction utilisé dans le choix des catégories peut être responsable de biais, en concluant à la nécessité d'un apport plus quantitatif. Ne disposant pas de système de capture du regard, nous avons dû annoter manuellement cette piste, tout en lançant en parallèle une étude sur l'annotation automatique de la direction du regard par traitement d'images, en cours de développement. En attendant que cet outil soit suffisamment performant, nous avons réalisé une annotation grossière, pour indiquer les segments temporels pendant les quels les yeux sont fermés (« close »), les segments où le regard est dirigé vers l'espace de signation placé devant le locuteur (« ssp ») et les segments où le regard n'est pas visible (mains devant le visage, tête baissée...). Mains : Les deux pistes dédiées aux mains ont été utilisées pour une étude en sciences du mouvement et ne sont pas décrites ici. Signes : Cette piste est utilisée pour identifier la catégorie de l'unité gestuelle, en distinguant les signes lexicaux, les structures iconiques, les pointages ou d'autres types d'unités (dactylologie, prise de rôle...). Il s'agit d'une annotation sujette à plus d'interprétation que les précédentes qui ne nécessitaient pas de connaître et comprendre la LSF pour être annotées. Des informations plus formelles sur la structure du signe et la relation entre les deux mains le cas échéant ont été adjointes à cette piste. Trois personnes différentes ont annotées ces trois types de pistes de manière indépendantes.

4. Analyse des relations temporelles entre segments d'annotation

Afin de pouvoir analyser les relations temporelles entre les segments d'annotation de la piste Regard et de la piste Signe, les annotations ont été exportées puis manipulées au sein d'un script R afin de pouvoir calculer les relations temporelles et extraire des statistiques. Les annotations sont constituées, pour chaque piste, d'une liste de segments dont on connaît les temps de début et de fin, ainsi que la valeur d'annotation. Ces annotations sont tout d'abord exportées sous forme de tableaux dans des fichiers cvs puis importées dans le script. Les données sont ensuite filtrées selon les phénomènes dont on veut étudier les relations. Dans notre cas, nous avons gardé les occurrences ayant la valeur « ssp » dans le tableau Regard et la valeur « TTF » dans le tableau Signe. Après filtrage, chaque tableau comporte de l'ordre de 2000 occurrences, ce qui nous garantit des calculs statistiques fiables. Nous avons utilisé l'algèbre des intervalles de Allen pour décrire les relations entre les segments des pistes étudiées. Cette algèbre définit treize relations, qui représentent toutes les relations possibles entre deux intervalles (figure 4). Pour chaque signe de type TTF, le script repère tous les éléments du tableau Regard dont les bornes temporelles sont proches et calcule la relation par comparaison des temps de début et de fin des segments. Nos résultats préliminaires, à vérifier sur l'ensemble du corpus, confirment que pour les constructions analysées, les TTF, un segment regard est en relation de précédence avec un segment TTF (figure 5). Une vérification manuelle des occurrences qui ne respectent pas cette règle a montré qu'il s'agit de cas pour lesquels les signes sont soit réalisés entre le locuteur et la caméra, à une hauteur telle qu'il n'est pas possible de savoir si le regard est dirigé vers la caméra ou vers l'espace de signation, soit réalisés de part et d'autre de la tête, ce qui fait qu'il est impossible pour le locuteur de regarder ses mains.

5. Conclusion et perspectives

La méthode présentée ici permet d'identifier et de quantifier les relations temporelles les plus fréquences entre des phénomènes annotés sur deux pistes différentes. Les résultats obtenus permettront de concevoir des traitements automatiques ciblés qui pourront à leur tour être utilisés par exemple pour l'aide à l'annotation de vidéos de LSF. Cependant, on observe généralement bien plus de deux composantes en action pour un phénomène donné. Ainsi, le regard n'est pas le seul élément non manuel qui peut accompagner les signes de type TTF. On observe aussi des activités sur différents éléments du visage (sourcils, plissement des yeux, bouche, joues...). Cette méthode doit maintenant être étendue afin de pouvoir traiter plus de deux pistes à la fois.

6. Figures

FIGURE 1 – Configuration des caméras infra-rouge et emplacement des marqueurs

FIGURE 2 – Exemples de photos de la première tâche

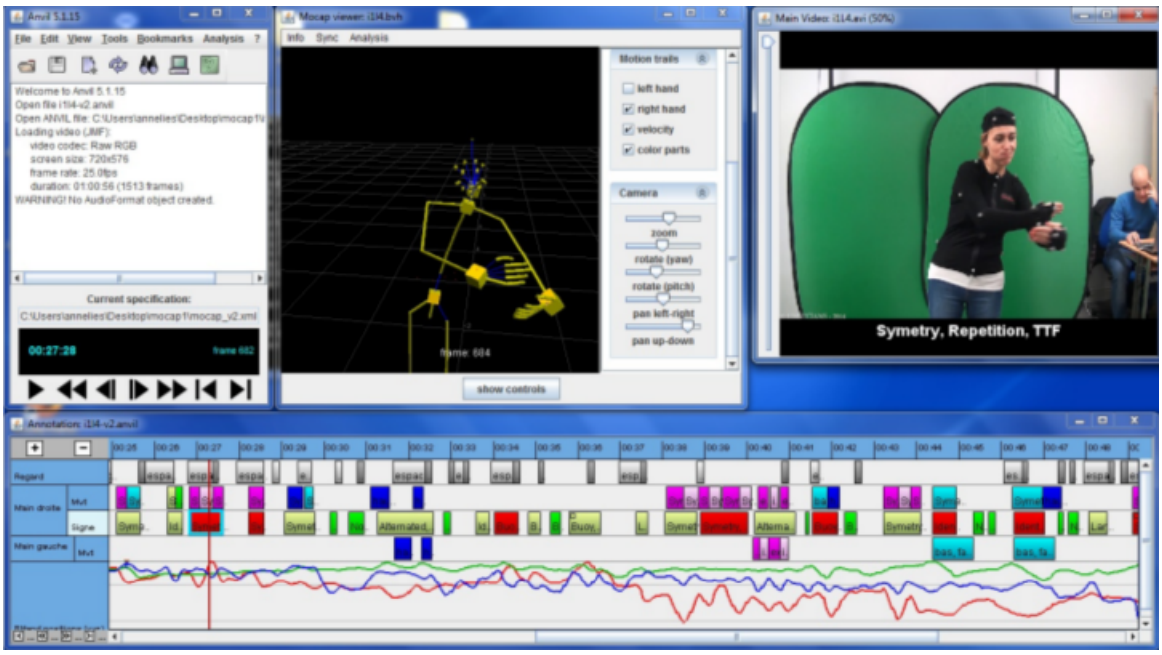


FIGURE 3 – Logiciel d'annotation ANVIL, avec visualisation des données de mocap

Relation	Illustration	Interpretation
$X < Y$ $Y > X$	X Y	X takes place before Y
$X\ \mathbf{m}\ Y$ $Y\ \mathbf{mi}\ X$	X Y	X meets Y (*i* stands for ***inverse***)
$X\ \mathbf{o}\ Y$ $Y\ \mathbf{oi}\ X$	X Y	X overlaps with Y
$X\ \mathbf{s}\ Y$ $Y\ \mathbf{si}\ X$	X Y	X starts Y
$X\ \mathbf{d}\ Y$ $Y\ \mathbf{di}\ X$	X Y	X during Y
$X\ \mathbf{f}\ Y$ $Y\ \mathbf{fi}\ X$	X Y	X finishes Y
$X = Y$	X Y	X is equal to Y

FIGURE 4 – Les 13 relations de Allen

FIGURE 5 – Les deux relations les plus fréquentes entre l'activité du regard et les TTF

Références bibliographiques

Benchiheub, M., Berret, B. and Braffort, A (2016). *Collecting and Analysing a Motion-Capture Corpus of French Sign Language*, 10th LREC Workshop on the Representation and Processing of Sign Languages : Corpus Mining, ELRA.

Braffort, A. (2014). *Eye gaze annotation practices : Description vs. Interpretation*, 6th LREC Workshop on the Representation and Processing of Sign Languages : Beyond the Manual Channel, ELRA

Cuxac, C. (2000). La langue des Signes Française (LSF) : les voies de l'iconicité, *Faits de Langues* n° 15-16, Ophrys, Paris.

Engberg-Pedersen, E. (1999). Eye gaze in Danish Sign Language monologues : Forms, functions, notation issues. *Intersign project papers*, v. 3, 33 -37.
http://www.sign-lang.uni-hamburg.de/intersign/workshop3/engbergpedersen.html

Heloir, A., Neff, M. and Kipp, M. (2010). *Exploiting Motion Capture for Virtual Human Animation : Data Collection and Annotation Visualization.* LREC Workshop on Multimodal Corpora : Advances in Capturing, Coding and Analyzing Multimodality, ELRA.

Johnston, T. (2014). Auslan Corpus Annotation Guidelines.
http://media.auslan.org.au/attachments/Johnston_AuslanCorpusAnnotationGuidelines_14June2014.pdf

Pragmatique du discours de lutte pour la justice climatique et énergétique : du corpus « maison » au corpus web

Camille Biros , Caroline Rossi et Inesa Sahakyan
ILCEA 4 – Univ. Grenoble Alpes
Prenom.Nom@univ-grenoble-alpes.fr,

1. Introduction

L'émergence d'un mouvement global de défense de la justice climatique et énergétique ne fait plus aucun doute, même s'il n'est pas toujours désigné en ces termes (UNEP, 2007 :314). Cependant, la reconnaissance officielle des préoccupations et revendications qui sont au cœur du mouvement est timide et tardive, et leur intégration au sein du discours sur le changement climatique semble toujours difficile. Ce constat est-il le signe d'une opposition toujours marquée entre le discours officiel et celui d'une ou plusieurs autres « communautés de discours » ? C'est ce que suggère le texte de l'accord de Paris, qui intègre la notion aux considérants et la rattache d'emblée à un ensemble de personnes (« certains ») qui y seraient particulièrement attachées :

> *Notant* qu'il importe de veiller à l'intégrité de tous les écosystèmes, y compris les océans, et à la protection de la biodiversité, reconnue par certaines cultures comme la Terre nourricière, et notant l'importance pour certains de la notion de « justice climatique », dans l'action menée face aux changements climatiques,

Pour tenter de rendre observable et de caractériser la variation des discours sur la justice climatique et énergétique, nous avons constitué deux types de corpus que nous présentons et analysons dans cette étude.

2. Méthode

Notre travail fait fond sur les acquis de la linguistique de corpus, mais en y intégrant un ingrédient fondamental de la pragmatique de corpus : la lecture horizontale d'un ensemble de textes assez homogènes (Rühlemann and Aijmer, 2014). Ces textes sont des rapports en anglais produits par des organisations onusiennes, des organisations non-gouvernementales environnementales et des entreprises du secteur énergétique, et ils constituent l'un des points d'ancrage des analyses proposées. A celles-ci s'ajoutent des analyses de corpus plus traditionnelles, basées sur la lecture verticale que permettent les concordanciers *Antconc* (Anthony, L. 2014) et *TXM* (Heiden, Magué, Pincemin 2010) et le logiciel d'étiquetage sémantique *WMatrix* (Rayson, 2008).
Les rapports d'organisations et programmes onusiens, des organisations non-gouvernementales environnementales et des entreprises du secteur énergétiques constituent trois « mini corpus maison » (traduction du terme anglais « DIY corpora », cf. Loock, 2016) de presque un million de mots pour chacun.
Une analyse des mots clés conduite dans *Antconc* fait ressortir les caractéristiques de ces rapports. Les différences sont exploitées pour aboutir à trois séries de termes que nous utilisons comme « semences » afin de fabriquer trois corpus web de taille comparable à nos corpus maison, en utilisant l'outil *BootCaT* (Baroni, S. and Bernardini, S., 2004). La constitution de sous-corpus différenciés avec l'outil *BootCaT* est rendue possible par un tri préalable des adresses web.
Nous utilisons une technique de visualisation simple (les nuages de mots) afin d'illustrer la plus grande diversité des données recueillies sur le web ainsi que la présence de « bruit » dans ce corpus. Cette dernière étape permet d'éliminer certaines sources non pertinentes, et de produire une liste cohérente des sites web utilisés pour constituer le corpus. Nous montrons que le recueil d'adresses web triées, en complément aux termes semences, permet à l'outil *BootCat* d'aboutir à des résultats plus pertinents pour cerner les caractéristiques lexicales et pragmatiques d'une communauté de discours.

3. Analyses

Pour analyser la pertinence de chaque sous-corpus pour éclairer le thème de la justice climatique, nous commençons par considérer la présence de ce terme et de ses synonymes. Si le thème est présent

dans les six sous-corpus, ses occurrences sont très peu nombreuses dans certains corpus, ce qui suggère qu'il appartient à un certain cadre idéologique.
Dans un deuxième temps, nous utilisons le logiciel d'étiquetage sémantique *WMatrix* pour comparer les grands thèmes abordés dans chaque sous-corpus. Nous repérons des variations signifiantes entre les trois sous-corpus maison, notamment concernant le thème de la responsabilité et de la vulnérabilité, que nous ne retrouvons pas dans les trois corpus web.
Dans un troisième temps, une analyse différenciée des mots composés avec le noyau « energy », grâce à la fonction « coocurrences » du logiciel *TXM*, permet de souligner des variations fines entre les trois sous-corpus maison concernant le traitement du thème de l'énergie. Nous observons dans quelle mesure ces variations se retrouvent dans les trois sous-corpus constitués sous *BootCat*.
Les résultats montrent que, malgré l'utilisation de mots clés différenciés, les différences plus fines repérées dans le corpus maison ne se retrouvent pas dans le corpus web. Ces résultats nous permettent d'interroger les différences de nature entre le corpus web et le corpus maison, notamment en lien avec la typologie des textes, puisque le corpus maison est constitué d'un genre de discours homogène, le rapport, alors que le corpus web présente une plus grande hétérogénéité générique.

4. Conclusion

Nos conclusions suggèrent que s'il existe une différence de nature entre corpus maison et corpus web, les données les plus riches se trouvent probablement toujours à la croisée des chemins. Nous plaidons donc pour l'usage conjugué et néanmoins différencié de corpus d'assez petite taille. Les étapes de constitution de ces corpus dessinent un cheminement méthodologique au terme duquel la spécificité de chaque corpus est soulignée, si bien que la somme des données utilisées ne se confond pas avec un seul, grand corpus. L'impact de telles analyses pour la constitution d'un corpus de langue de spécialité environnementale [1], qui est à l'arrière plan de la présente étude, sera discuté en fin de parcours.

Références bibliographiques

Anthony, L. (2014). AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan : Waseda University. Available from http://www.laurenceanthony.net/

Austin, J.L. (1975). *How to Do Things with Words.* Cambridge, Mass. : Harvard University Press.

Baroni, M. & Bernardini, S. (2004). BootCaT : Bootstrapping Corpora and Terms from the Web. In *LREC*. http://www.cs.utah.edu/nlp/readinglist/BaroniB04.pdf

Garcia, P. & Dresher, N. (2006). *Corpus-Based Analysis of Pragmatic Meaning, in Corpus Linguistics : Applications for the Study of English.* Bern : Peter Lang.

Heiden, S., Magué, J-P., Pincemin, B. (2010). *TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement.* In Sergio Bolasco, Isabella Chiari, Luca Giuliano (Ed.), Proc. of 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010) (Vol. 2, p. 1021-1032). Edizioni Universitarie di Lettere Economia Diritto, Roma, Italy.

Loock, R. (2016). L'utilisation des corpus électroniques pour le traducteur professionnel. Bruxelles, 9 juin 2016, DGT.

Maranville, Angela R., Tih-Fen Ting, et Yang Zhang. (2009). An environmental justice analysis : superfund sites and surrounding communities in Illinois. *Environmental Justice* 2 (2) : 49-58.

Rayson, P, 2008, From key words to key semantic domains. *International Journal of Corpus Linguistics.* 13 :4 pp. 519-549. DOI : 10.1075/ijcl.13.4.06ray.

Rühlemann, C. & Aijmer, K. (2014). Corpus pragmatics : laying the foundations, in *Corpus Pragmatics. A Handbook.* Aijmer, K. and Rühlemann, C. (eds.), Cambridge : Cambridge University Press.

United Nations Environment Programme, éd. (2007). *Global environment outlook : environment for development, GEO 4.*, Nairobi, Kenya : London : United Nations Environment Programme ; Stationery Office [distributor].

1. Nous remercions le consortium CORLI qui a financé les dernières étapes de la préparation de ce corpus. Le corpus sera bientôt partagé sur Ortolang.

Session 1.B.
Analyse du discours et des textes

Analyse du discours, linguistique de corpus et données numériques : quelle rencontre ? A propos de la locution « dans DET cadre »

Emilie Née et Frédérique Sitri
UPC, Ceditec
Université Paris Nanterre, Modyco
nee.emilie@gmail.com, fsitri@u-paris10.fr

Nous souhaitons ici contribuer à la réflexion sur les conditions de possibilité et les modalités d'une rencontre entre analyse de discours « outillée » et linguistique de corpus, à travers l'exemple de l'étude de la locution « dans DET cadre ».

L'analyse de discours, qui vise à rendre compte des déterminations de tous ordres qui pèsent sur la matérialité discursive, s'intéresse tout particulièrement depuis une vingtaine d'années à l'articulation entre les formes ou configurations langagières et les contraintes liées aux genres de discours, ces derniers étant envisagés à la suite de Bakhtine (1984) dans leur relation à une sphère d'activité. Les travaux qui se réclament de la linguistique de corpus quant à eux visent à rendre compte des usages de la langue tels qu'ils peuvent être saisis dans des corpus rendus disponibles par la numérisation des données et leur accessibilité via internet, en accordant une attention particulière à la variation de ces usages. On peut donc considérer que l'on a affaire à des cadres théoriques différents qui définissent des cadres d'analyses distincts avec des incidences en termes de constitution de corpus. Cependant, le recours à une exploration outillée en analyse de discours tout comme l'attention grandissante portée aux conditions de production des données analysées en linguistique de corpus favorise l'interaction entre les deux approches. Quel est l'impact du partage accru de données numériques (corpus en open access, bases de données ou « réservoir de corpus ») sur la conception du corpus en analyse de discours outillée (passage de corpus clos, homogènes à des corpus ouverts, en dialogue / interaction) ? Comment articuler déterminations génériques et prise en compte des usages et de leur variation en utilisant les ressources textuelles (disponibles) ? Telles sont les questions que nous soulèverons en prenant l'exemple d'une étude en cours sur la locution « dans DET cadre ».

Dans le corpus de rapports éducatifs sur lesquels nous travaillons depuis un certain temps, la locution « dans DET cadre » est apparue remarquable par sa fréquence et sa saillance, et nous avons pu l'analyser comme une « routine » discursive (Née, Sitri, Veniard 2016). Pour décrire les particularités de son emploi dans le genre « rapport éducatif », nous avons élargi notre enquête à des corpus diversifiés, construits ou « issus du web » (Googlebooks).

Dans une première étape de cette recherche, nous avions mis en parallèle les emplois de « dans DET cadre » dans le corpus de rapports éducatifs et ses emplois dans un corpus de dépêches de presse (réuni par Marine Damiani) d'une part et d'articles scientifiques (corpus Scientext, Tutin et Grossmann 2014) d'autre part, en faisant porter nos observations sur le niveau syntaxique (place de la locution, degré d'intégration syntaxique, nature du déterminant) et sur le plan sémantique (nature du « cadre »).

Cette première approche nous a permis de faire apparaître que la séquence « dans DET cadre », si elle est relativement figée en langue, donne lieu à des usages diversifiés selon les genres de discours (Née et Sitri, 2014). Ainsi, dans les rapports éducatifs, la locution a rarement une fonction cadrative, et renvoie le plus souvent, sur le plan sémantique, au cadre institutionnel et légal de l'intervention éducative (type de placement, type de mesure) ou aux conditions de l'intervention. Plus largement, on a proposé de mettre en relation l'emploi de la locution dans les rapports éducatifs avec l'importance du « cadre » dans le discours des éducateurs (Cambon 2006) – cadre qu'ils sont censés imposer aux personnes dont ils s'occupent et cadre règlementaire dans lequel s'insère leur action.

Poursuivant le travail amorcé, nous nous nous proposons d'affiner nos analyses d'une part en explorant d'autres genres et sphères d'activité mais aussi la variation écrit/oral (corpus ESLO et CFPP2000) voire la variation diachronique (Corpus ESLO 1 et 2, Frantext, Googlebooks via Ngram viewer).

Notre objectif est ainsi double. D'une part il s'agit de poursuivre la mise en évidence des spécificités d'emploi de la locution dans le corpus princeps. A ce titre, la référence fréquente au cadre légal de l'intervention nous conduit à explorer un corpus de textes législatifs et réglementaires.

D'autre part, dans la lignée de Sitri et Veniard 2016, nous souhaitons corréler ces emplois avec la mise en évidence d'une « montée en puissance » de la locution, particulièrement dans des genres écrits ou dans des genres oraux influencés par l'écrit, ce que devrait permettre à la fois la prise en compte de corpus oraux et la perspective diachronique. Le corpus ESLO permet ainsi de corréler l'emploi de la locution avec le statut du locuteur, le genre de l'interaction, d'établir s'il constitue une reprise, et de se pencher sur d'éventuelles évolutions en termes de fréquence ou d'usage entre ESLO 1 et ESLO2.

D'un point de vue interprétatif, et du côté de l'analyse du discours, on interrogera aussi la montée en puissance de la locution et son usage marqué dans certains genres avec la place de la notion de cadre dans les sphères d'où ils émanent.

On voit que la mise en œuvre de la rencontre entre analyse du discours et linguistique de corpus rencontre la réflexion en cours depuis un certain temps sur la question de l'hétérogénéité des corpus : ici, la mise en évidence, dans un corpus d'observation largement documenté sur le plan de ses conditions de production, d'une forme saillante que nous considérons comme une « routine », conduit à explorer les usages de cette même forme dans d'autres corpus conçus comme corpus « de comparaison ».

Références bibliographiques :

Bakhtine M., (1984 [1952-1953]). *Les genres du discours* in *Esthétique de la création verbale*. Paris : Gallimard.

Brunet, E., Vanni, L. (2014). « GOOFRE version 2 », *Actes des JADT2014*, p. 105-119, URL : http://www.jadt.org/.

Gadet F., Wachs S. (2015). « Comparer des données de corpus : évidence, illusion ou construction ? ». *Langage et société*, 154, p. 33-49.

Benzitoun, C. (2014). « La place de l'adjectif épithète en français : ce que nous apprennent les corpus oraux », SHS Web of Conferences 8 (2014), DOI 10.1051/shsconf/20140801066

Cambon, L. (2009). *L'identité professionnelle des éducateurs spécialisés. Une approche par les langages.* Rennes : Presses de l'École des Hautes Études en Santé Publique.

Fagard, B. et Combettes, B. (2013). « De en à dans, un simple remplacement ? Une étude diachronique ». *Langue française*, 178, p. 93-115. DOI : 10.3917/lf.178.0093.

Marcon, M (2012). Mieux vaut n-gram qu'introspection. Google Ngram Viewer et parémiologie. *Paremia*, Asociación Cultural Independiente, 2012, p. 85-95.

Née, É., Sitri, F. et Veniard M. (2016). « Les routines, une catégorie pour l'analyse de discours : le cas des rapports éducatifs », *Lidil*, 53, p. 71-93.

Née, É. et Sitri, F. (2014). « « Dans det cadre (+ spécification)» : de la locution à la routine discursive », communication au congrès WRAB. Paris Nanterre.

Sarda L. (2010). « Les adverbiaux prépositionnels en *dans* : exploration en corpus de la notion de contenance », *Corela* [En ligne], HS-7 | 2010, mis en ligne le 31 mai 2010, consulté le 29 janvier 2017. URL : http://corela.revues.org/911~ ; DOI : 10.4000/corela.911

Sarda, L. et Carter-Thomas, S. (2012). « L'impact de la position phrastique sur les fonctions et valeurs des SP adverbiaux : l'exemple des SP en sur et dans », *Travaux de linguistique*, 64, p. 21-54. URL : http://www.cairn.info.faraway.u-paris10.fr/revue-travaux-de-linguistique-2012-1-page-21.htm DOI : 10.3917/tl.064.0021

Tutin, A.et Grossmann, F. (eds) (2014). *L'écrit scientifique : du lexique au discours. Autour de Scientext*, Presses de l'Université de Rennes.

Beware of negative utterance; they can be most devastating. The Use of Negation on Twitter during the 2014 European Elections

Elena Albu
Université de Strasbourg, France
elenaoac@gmail.com

Negation is a heterogeneous phenomenon leading to multiple meanings and raising different problems of interpretation. While much has been said about how negation is produced and interpreted in oral communication (Ducrot 1972, 1984; Horn 2001; Kaup et al. 2007; Moeschler 1993, 2013, to name but a few), little is known about how negation is employed in computermediated communication (Longhi et al. 2016). This paper aims at discussing the use of negation on the micro-blogging service Twitter during the European Elections (EE) held in May 2014. More specifically, particular attention will be paid to the following research questions:

1. What types of negatives utterances are used in the electoral tweets during the 2014 EE?
2. To what extent does negation influence the production of an electoral tweet and shape the candidates' political discourse on Twitter?

The first research question focuses on the characterization of the negative utterances used on Twitter in the 2014EE. Building on the classification given in Longhi et al. (2016), the emphasis will be placed on identifying how many types of negative utterances are used and on describing their main features based on syntactic, semantic and pragmatic criteria. The position of negation within the tweet and whether the negative clause is used alone or in combination with different clauses meant to correct, contrast or add more information will also be investigated.

The second research question aims at discussing the argumentative force of negation and the impact negation has on the political discourse in the institutional context of 2014 EE. The relation between different negative utterances and their discursive functions will be discussed from a genre perspective, on the one hand, and the extent to which negation plays a role in the political discourse in the present institutional context will be analyzed, on the other hand. The following questions will be of particular interest:

(a) Does negation have a descriptive function or is it used to refute or contradict an opinion previously stated or implied in a preceding tweet?

(b) Is the tweet as a discursive genre used in connection with a particular type of negative utterance during the 2014 EE?

The corpus of data under investigation in this paper is exclusively represented by the tweets of the UK candidates collected within a time span of two weeks prior to the election day. In order to answer the above-mentioned research questions, both a corpus-based and corpus-driven analysis will be carried out. The software used for concordancing and text analysis is represented by AntConc (Anthony 2014).

This investigation is part of the international project: *Twitter at the European Elections: A Comparative International Study of the Use of Twitter by Candidates at the European Parliamentary Elections in May 2014* [1], run by the Human Sciences Institute of Dijon.

Key words: negation, computer-mediated communication, Twitter, genre, corpus linguistics, discourse analysis, political discourse

Références bibliographiques

Anthony, L. (2014): AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/.

1. http://iutdijon.u-bourgogne.fr/pedago/src/politicsmediastic/en/tee2014-2/

Ducrot, O. (1972): *Dire et ne pas dire. Principes de sémantique linguistique*, Paris: Hermann.

Ducrot, O. (1984): *Le Dire et Le Dit*, Paris: Les Éditions de Minuit.

Horn, L. (2001): *A natural history of negation*, University of Chicago Press.

Longhi, J. et al. (2016): Extraction automatique de phénomènes linguistiques dans un corpus de tweets politiques : quelques éléments méthodologiques et applicatifs à propos de la négation, in Silvia Palma, Emilia Hilgert, Rene Daval, Pierre Frath. *Res per Nomen V*, Presses universitaires de Reims.

Kaup et al. (2007): Experiential simulation of negated text information. *Quarterly Journal of Experimental Psychology*, 60, p. 976-990.

Moeschler, J. (1993): « Une, deux ou trois négations? », *Langue Française*, p. 94, 8-25.

Moeschler, J. (2013): 'How 'Logical' are Logical Words? Negation and its Descriptive vs. Metalinguistic Uses', in Taboada M. & Trnavac R. (eds.), *Nonveridicality, evaluation and coherence relations*, Leiden, Brill, p. 76-110.

Session 2.A.
Transcription de l'oral

Agrégation automatisée de corpus de français parlé

Christophe Parisse [1], Christophe Benzitoun [2], Carole Etienne [3] et Loïc Liégeois [4]
[1]Inserm, Modyco, Université Paris Nanterre
cparisse@u-paris10.fr
[2]Atilf, Université de Lorraine
[3]Icar, CNRS
[4]CLILLAC-ARP et LLF, Université Paris Diderot

1. La question des données en français parlé

Les recherches en linguistique de l'oral et en psycholinguistique demandent de plus en plus l'accès à de larges corpus diversifiés. Dans ce cadre, on utilise soit un corpus élaboré spécifiquement pour répondre à un objectif (PFC, ESLO, CFPP) ou à une thématique (TCOF, CLAPI), soit des bases de données ou de corpus organisées et orientées autour de projets ou de questions plus générales (FLORAL, ORFEO). Si ces approches présentent l'avantage de permettre de travailler à partir d'un matériel contrôlé, uniformisé et de qualité garantissant la fiabilité des études, elles se heurtent toutefois au coût et à la durée de création et de traitement des corpus. Cette présentation propose d'explorer une troisième voie : celle de l'agrégation automatisée de corpus préexistants, dont nous allons exposer les différentes étapes.

La création de nouvelles ressources orales est particulièrement coûteuse. Une des particularités des corpus oraux est qu'ils nécessitent une transcription, qui est en définitive une interprétation d'une forme orale ou multimodale (Morgenstern & Parisse, 2007). Ceci signifie que la donnée primaire d'un corpus oral doit être l'enregistrement audio ou vidéo, qui permet de vérifier, remettre en cause, modifier ou compléter les transcriptions et surtout l'interprétation qui est faite de la source initiale. Un corpus oral doit donc s'accompagner de médias auxquels s'ajoutent une ou plusieurs transcriptions manuelles au coût de réalisation important.

Lorsque la constitution de nouveaux corpus ou la création d'une nouvelle ressource homogène et uniformisée à partir de corpus déjà existants n'est pas possible, par manque de moyens, temps ou financement, on peut agréger des données déjà disponibles. Il s'agit de constituer dynamiquement des collections de données à partir de sources multiples. Le principe est de réunir, dans une seule base, des données de corpus disponibles provenant de sources hétérogènes. Cette technique présente des avantages au niveau de la rapidité, de la variété, du volume et de l'évolutivité : toute source nouvelle ou toute source ancienne mise à jour permet d'augmenter en temps réel la quantité de données disponibles. En revanche, une telle collection regroupant à la volée des données provenant de corpus décrits de différentes manières, annotant différents phénomènes de l'oral et dans différentes conventions de transcription invite à la prudence lors de leur interrogation, car l'homogénéité des objets manipulés peut être de mauvaise qualité. De plus, la représentativité de chaque genre ne sera pas forcément respectée compte-tenu des différences de volume et de nature des sources utilisées.

C'est la raison pour laquelle un projet comme ORFEO, dont le but est de construire un Corpus d'Etude pour le Français Contemporain, a fait le choix d'uniformiser les conventions de transcription et de diversifier les situations de parole. Mais cela représente un coût financier non négligeable et nécessite un temps de réalisation conséquent. Nous nous proposons ici de présenter une autre méthode pour constituer rapidement une large collection de données orales sans nécessairement viser une homogénéisation parfaite des sources qui la composent, mais en nous inspirant des acquis de projets comme ORFEO dans le traitement des transcriptions et des métadonnées.

En fin de présentation, nous effectuerons une analyse linguistique pour illustrer les potentialités offertes par notre méthodologie de constitution d'une collection de corpus hétérogènes.

2. Format TEI-CORPO - Tei Oral ISO

Pendant de nombreuses années, la trop grande hétérogénéité des données et le faible nombre de corpus diffusés ont empêché un quelconque regroupement des données à des fins d'analyse linguistique. Aujourd'hui, l'arrivée dans le paysage de la recherche de consortiums de linguistique comme CORLI (cf. http://www.huma-num.fr/consortiums/) pour consolider et diffuser de bonnes pratiques dans la

continuité des travaux déjà entrepris (Baude & al., 2006), permet de tester la faisabilité et l'apport d'une collection de corpus existants. Cela a déjà été tenté dans des travaux antérieurs (Benzitoun & Bérard, 2010), mais avec de plus grandes difficultés qu'aujourd'hui : absence d'outils de conversion, hétérogénéité des symboles utilisés (un même symbole pouvant avoir une sémantique différente), trucages orthographiques, problèmes d'encodage de caractères provenant des traitements de texte utilisés, etc.

La large diffusion du format TEI dans les publications de corpus en linguistique a amené le consortium IRCOM à proposer l'utilisation d'un standard TEI (ISO 24624) décrivant son utilisation pour l'oral. Le consortium CORLI (anciennement IRCOM) et l'Equipex Ortolang ont mis en commun leurs ressources pour créer un logiciel de conversion TEI-CORPO (Liégeois et al., sous presse) permettant la conversion en TEI des formats des logiciels de transcription de l'oral comme CLAN, ELAN, Praat ou Transcriber. Cet outil permet d'utiliser la TEI comme un format pivot entre les logiciels de transcription, mais également comme un format commun sur lequel s'appuyer pour assembler des corpus d'origines différentes ou effectuer de nouveaux traitements.

3. Une collection de français parlé de plus de 8 millions de mots

Nous proposons de créer une collection dynamique à partir de différents corpus sous condition qu'ils soient librement disponibles pour la recherche et disposent d'un matériel audio ou vidéo associé. Nous utilisons pour cela le logiciel de conversion que nous améliorons en fonction des besoins ou des difficultés rencontrées. Les améliorations portent principalement sur trois points :

- inclusion de nouvelles métadonnées, propagation de métadonnées existantes (données sur la situation et sur les caractéristiques des locuteurs) ;
- génération d'une variante uniformisée des formats des transcriptions ;
- application d'une analyse morphosyntaxique paramétrable.

Ces améliorations nous permettront de mener notre analyse linguistique de la collection de corpus constituée en prenant comme variables d'étude l'âge des locuteurs, la situation d'interaction et/ou le corpus d'origine.

L'application pratique a permis d'intégrer des données des corpus suivants, disponibles sur CHILDES (Lyon), Cocoon (Eslo), Ortolang (Alipe, Colaje, Mpf, Pfc, Tcof, Clapi), Cfpp2000 (http ://cfpp2000.univ-paris3.fr/). Ces corpus sont disponibles dans les formats CLAN, Praat et Transcriber (voir références).

On obtient un total de 8,5 millions de mots, hors ponctuation. La mise en œuvre pose des problèmes d'uniformisation des transcriptions en fonction des formats d'origine. Certains formats sont plus normalisés, comme les transcriptions CLAN, tandis que d'autres sont plus variables (Praat). Nous proposons, plutôt que d'éditer les transcriptions d'origine, de fournir des procédures d'uniformisation des formats éventuellement adaptées à un cas particulier. Ces procédures permettent de garder le caractère dynamique de cette grande collection de données. Elles seront intégrées à l'outil de conversion.

L'existence d'un format unique facilite également la mise en œuvre de procédures variées pour l'analyse des données. Il est ainsi possible de faire varier l'outil d'étiquetage en parties du discours, la lemmatisation ou la tokenisation des données en utilisant des outils open-source. L'ensemble des logiciels utilisés est disponible sur le site thématique langage oral d'Ortolang (http://ct3.ortolang.fr/tei-corpo/).

4. Conclusion

Nous avons démontré la faisabilité de la création dynamique et ajustable d'une vaste collection de corpus de langage oral par le partage de formats communs de données et par la mise en œuvre de bonnes pratiques dans la création et la diffusion de corpus. Une telle collection va cumuler à la fois les qualités et les défauts des corpus existants. Si la solution que nous proposons ici peut être adaptée pour répondre à un certain nombre de questions linguistiques, nous sommes conscients des limites qu'elle présente :

- L'hétérogénéité des annotations et des conventions de transcription va poser des problèmes pour la formulation de requêtes et l'utilisation de logiciels d'annotation automatique tolérant généralement mal les variations graphiques. Le travail de correction, même automatisé, peut alors s'avérer coûteux et diminuer l'intérêt de la procédure.
- La représentativité d'une collection correspondant aux corpus « disponibles » pour le partage ne peut concurrencer celle de projets organisés.

Toutefois, ces limites ne doivent pas empêcher le développement de travaux visant à l'agrégation de corpus alors que certaines approches linguistiques cherchent de plus en plus à tester leurs hypothèses sur des ensembles toujours plus vastes de données orales. Au contraire, les problématiques liées à l'hétérogénéité des données sources représentent un défi technique qui requiert, au niveau linguistique, une profonde réflexion sur ce que pourrait être un niveau commun de représentation des données primaires, c'est à dire un niveau commun de transcription. Dans cet objectif, il sera possible dans un futur proche de s'appuyer sur des projets comme ORFEO qui, en plus de regrouper des corpus déjà constitués, ont corrigé, complété et uniformisé les transcriptions et les métadonnées. On peut aussi espérer que la généralisation de la pratique du partage et de la diffusion des corpus et de leurs outils de traitement permettra à terme de réunir la large communauté de la linguistique orale et multimodale autour du partage des annotations et des conventions acceptées par tous, comme ce fût le cas au sein de la communauté des acquisitionnistes dans le projet CHILDES.

Références bibliographiques

ALIPE : Liégeois, L., Chanier, T. et Chabanal, D. (2014). *Corpus globaux ALIPE : Interactions parents-enfant annotées pour l'étude de la liaison.* Nancy : Ortolang. [http://hdl.handle.net/11041/alipe-000853]

Baude O., Blanche-Benveniste C., Calas M.-F., Cappeau P., Cordereix P., et al. (2006). *Corpus oraux, guide des bonnes pratiques.* CNRS Editions, Presses Universitaires Orléans, pp.203.

Benzitoun, C. & Bérard, L. (2010), Mutualisation et uniformisation de ressources de français parlé, in Azzopardi S. (coord), *Corpus, Données, Modèles, Cahiers de Praxématique,* 54-55, PULM : Montpellier, pp. 175-188, paru en 2013.

CFPP2000 : Branca-Rosoff S., Fleury S., Lefeuvre F., Pires M., (2012). "*Discours sur la ville. Corpus de Français Parlé Parisien des années 2000 (CFPP2000)*".

Clapi : Groupe ICOR (H. Baldauf-Quilliatre, I. Colon de Carvajal, C. Etienne, E. Jouin-Chardon, S. Teston-Bonnard, V. Traverso) (2016)," CLAPI, une base de données multimodale pour la parole en interaction : apports et dilemmes ", In Avanzi M., Béguelin M.-J. & Diémoz F. (eds), Corpus de français parlés et français parlés des corpus, Cahiers Corpus.

COLAJE : Morgenstern, A. & Parisse, C. (2012). The Paris Corpus. *Journal of French Language Studies*, 22, 7-12.

ESLO : Eshkol-Taravella I., Baude O., Maurel D., Hriba L., Dugua C., Tellier I., (2012). Un grand corpus oral « disponible » : le corpus d'Orléans 1968-2012. in *Ressources linguistiques libres*, TAL. Volume 52 – n/upo 3/2011, 17-46.

ISO 24624 Language resource management - Transcription of spoken language - https://www.iso.org/obp/ui/#!iso:std:37338:en

Liégeois, L., Etienne, C., Benzitoun, C., Chanard, C. et Parisse, C. (sous presse). Using the TEI as pivot format for oral and multimodal language corpora. *Journal of the Text Encoding Initiative*, 10.

LYON : Demuth, K. & A. Tremblay (2008). Prosodically-conditioned variability in children's production of French determiners. *Journal of Child Language*, 35, 99-127.

Morgenstern, A. et Parisse, C. (2007). Codage et interprétation du langage spontané d'enfants de 1 à 3 ans. *Corpus*, 6, p. 55-78.

MPF : Gadet F. (2013), Collecting a new corpus in the Paris area : Intertwining Methodological and Sociolinguistic Reflections, in D. Hornsby, M. Jones (Eds), *Language and Social Structure in Urban France*. Oxford : Legenda, 162-171.

PFC : Durand, Jacques, Bernard Laks & Chantal Lyche (2009). Le projet PFC : une source de données primaires structurées. In J. Durand, B. Laks et C. Lyche (eds) *Phonologie, variation et accents du français*. Paris : Hermès. pp. 19-61.

TCOF : Analyse et traitement informatique de la langue française - UMR 7118 (ATILF) (2017). TCOF : Traitement de Corpus Oraux en Français [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, https://hdl.handle.net/11403/tcof/v1.

Le transcripteur transcrit : retour d'expérience à partir du corpus des ESLO

Linda Hriba , Olivier Baude et Céline Dugua
LLL UMR 7270, Université d'Orléans
MoDyCo UMR 7114, Université Paris Nanterre
linda.hriba@yahoo.fr

1. Les enquêtes sociolinguistiques à Orléans : un très grand corpus variationniste

Les Enquêtes sociolinguistiques à Orléans (ESLO) forment un grand corpus oral constitué de deux enquêtes réalisées à deux périodes distinctes. La première enquête ESLO1 (1968-1971) est un corpus clos de 470 enregistrements, soit 318 heures d'enregistrements qui représente – selon l'estimation de l'époque – 4,5 millions de mots. La seconde enquête (ESLO2), commencée au début des années 2000 et toujours en cours de réalisation, affiche un objectif de plus de six millions de mots pour 450 heures d'enregistrements.

ESLO ne constitue pas seulement un corpus de masse de données, il s'agit d'un réservoir de corpus conçu dans un souci de représentativité des pratiques linguistiques d'une communauté d'auditeurs dans une ville donnée et à des moments distincts. La prise en compte de la variation, et de toutes les variations est au cœur du projet et guide à la fois les choix méthodologiques qui ont été réalisés dès les premières étapes de la constitution du corpus, les regards que nous porterons sur les analyses, et également la question de la transcription.

2. Procédure de transcription des ESLO

Depuis 2003, le LLL (Orléans) s'est donné pour objectif de transcrire et rendre disponible l'intégralité du corpus ESLO. Face à l'ampleur de la tâche et soucieux de rendre rapidement accessible le corpus, la transcription repose sur des conventions minimales. Il s'agit de répondre à un simple objectif de navigation dans le corpus. Ces premières contraintes nous ont, par ailleurs, orientés vers le logiciel de transcription *Transcriber* qui permet de réaliser les alignements/synchronisations sontranscription très facilement et qui, grâce à son interface simple, constitue le meilleur outil pour réaliser des transcriptions « au kilomètre ».

Afin de définir nos conventions de transcription, nous avions entrepris une comparaison des pratiques au sein de grands projets (CLAPI, DELIC etc.) travaillant sur l'oral, ce qui avait permis la mise en évidence des principes généralement partagés par tous ces projets, principes que nous avons adoptés pour ESLO. Ces derniers reposent sur une transcription orthographique standard qui rend compte des phénomènes spécifiques de l'oral (répétitions, amorces etc.), avec une segmentation en tours de parole (les choix opérés sont disponibles dans le « Guide du transcripteur » sur le site ESLO). Ainsi pour toute analyse ultérieure, une reprise de la transcription avec des conventions répondant aux cadres théoriques du chercheur et/ou des niveaux d'annotations sont indispensables.

Le LLL a également porté une attention particulière aux travaux qui, depuis 1970, ont montré la nécessité de relire à plusieurs reprises les transcriptions Fillol et Mouchon (1977). La relecture n'est efficace que si elle est réalisée par une autre personne que le transcripteur (Lahire, 1981). En partant de ce constat, le LLL a donc décidé de recourir à trois « écouteurs » (Blanche-Benveniste & Jeanjean, 1987) distincts pour la transcription des ESLO. La transcription se fait alors en trois étapes successives :

- une version A (VA) qui correspond à une première version « brute » de transcription, la priorité est donnée à la synchronisation de la transcription avec l'enregistrement,
- une version B (VB) dans laquelle il s'agit de vérifier l'orthographe et le respect des conventions de transcription de la version A,
- et une version C (VC) qui est la relecture de la version B.

On obtient ainsi trois versions de transcription pour un même enregistrement avec trois transcripteurs différents.

Ainsi, la transcription n'est plus conçue uniquement comme le préalable à une étude sur corpus oraux, elle est une façon de mettre en perspective les conditions de productions des données. En ce sens, elle constitue une étape qui reflète le champ de la linguistique, les théories, l'inscription du chercheur dans son domaine, et également les « attitudes » et les représentations des transcripteurs.

3. Les transcriptions : un observatoire des variations

Cette procédure, qui consiste à disposer pour chaque enregistrement de trois versions de transcription, nous a permis de relever d'importantes divergences entre ces étapes. En nous appuyant sur un corpus constitué de 20 enregistrements (60 fichiers de transcription), nous avons, à l'aide d'un logiciel spécialisé *Beyond Compare*, comparé les différentes versions obtenues en les confrontant deux à deux (VA *vs* VB et VB *vs* VC). L'extraction et l'analyse des différences ont révélé l'impossible stabilisation d'une version définitive de transcription. Il s'avère qu'en moyenne, 330 interventions sont nécessaires pour passer d'une VA à une VC ; ces dernières concernent trois grandes catégories de variations :

- des variations graphiques. Elles regroupent l'ensemble des erreurs qui, d'une part, correspondent aux fautes induites par les outils de saisie (clavier) et d'aide à la transcription (*Transcriber*) et qui, d'autre part, sont le reflet d'un non-respect d'une norme orthographique ou de codage (cf. conventions de transcription propres au projet) ;
- des variations de segmentation. Elles concernent l'alignement temporel, la segmentation en sections (ils s'agit de types de questions ou de thématiques), en tours de parole ainsi que les pauses ;
- et des variations de perception manifestant des divergences d'écoute.

Différents paramètres acoustiques, linguistiques et sociologiques permettent d'expliquer ces variations de perception, dont en premier lieu les caractéristiques propres des transcripteurs. Sur ce dernier point, nous manquons d'informations sur nos transcripteurs et de données qui permettraient de mieux comprendre qui ils sont et quel est leur rapport à l'écrit.

4. Un nouveau module ESLO : entretiens avec les transcripteurs

Ce nouveau module, en cours de construction, consiste en des entretiens semi-directifs auprès des transcripteurs avec pour objectif de capter leurs représentations de la langue, de les situer au sein de pratiques sociales et de confronter ces éléments aux variations de transcription relevées dans le corpus. Ces entretiens s'organiseront en cinq grandes thématiques.

- la trajectoire des transcripteurs : depuis la période scolaire, puis universitaire jusqu'à leur activité professionnelle actuelle ;
- l'expérience de transcription : l'objectif sera de recueillir leurs ressentis sur cette activité, de faire émerger ce qui était le plus difficile et pénible mais aussi le plus plaisant ;
- des questions autour de leur rapport à la norme et à l'écrit, notamment à l'orthographe. Il s'agira par exemple de leur demander comment ils abordent l'écrit dans des nouveaux moyens de communication ;
- une thématique sur leur rapport à la lecture aujourd'hui, mais aussi durant leur apprentissage, au collège/lycée et à l'Université ;
- et enfin, des questions sur leurs pratiques d'écriture, et leurs pratiques culturelles.

L'analyse de ces entretiens se réalisera à trois niveaux. Nous souhaitons établir une échelle qui prenne en compte ces différents paramètres et qui éventuellement les pondère afin de proposer, pour chaque transcripteur, une catégorisation de ses pratiques et représentations. Nous réaliserons ensuite

une analyse du contenu des entretiens et enfin nous utiliserons ces éléments dans l'analyse des variations de transcription de chacun des transcripteurs.

Cette approche réflexive de la phase de transcription du corpus des ESLO souhaite dépasser une simple description de la méthodologie de linguistique de corpus oraux pour atteindre une analyse linguistique fondée sur des données attestées et situées. La transcription d'un très grand corpus oral, pour peu que celui-ci soit documenté par suffisamment d'éléments permettant de situer socialement la parole captée, offre incontestablement un point de vue privilégié sur le passage de la perception acoustique à l'empreinte linguistique.

Références bibliographiques

Barras C., Adda, G., Adda-Decker, M., Habert, B., Boula de Mareüil, P, Paroubek, P (2004). Automatic audio and manual transcripts alignments, time-code transfer and selection of exact transcripts. *Actes de la Fourth International Conference on Language Resources and Evaluation (LREC)*, Lisboa, May 2004, vol. 3, pp 877-880.

Baude, O. Dugua C. (2011). (Re)faire le corpus d'Orléans quanrante ans après : quoi de neuf, linguiste ? *Corpus, 10* Varia, 99-118.

Baude, O. (2006). *Corpus oraux, Guide des bonnes pratiques.* Paris, CNRS Editions.

Bilger, M. (ed) (2008). *Données orales, les enjeux de la transcription.* Les cahiers de l'Université de Perpignan.

Blanche-Benveniste, C., Jeanjean, C. (1987). *Le français parlé. Transcription et édition.* Paris, Inalf, Didier érudition.

Cappeau P., Gadet F. (2013). Quand l'œil écoute : que donnent à lire les transcriptions d'oral ? *Communication orale au CILPR,* Nancy.

Cappeau, P., Gadet, F., Guerin, E., Paternostro, R. (2011) « Les incidences de quelques aspects de la transcription outillée », *Linx* [En ligne], *64-65*, 85-100. Mis en ligne le 01 juillet 2014, consulté le 17 janvier 2017. URL : http://linx.revues.org/1403 ; DOI : 10.4000/linx.1403

Cappeau, P., Gadet, F., (2010). Transcrire, ponctuer, découper l'oral : bien plus que de simples choix techniques. *Cahiers de linguistique, 35/1*, 187-202.

Delais-Roussarie, E. & Yoon, H.-Y. (2011). Transcrire la prosodie : un préalable à l'échange et à l'analyse des données. *Journal of French Language Studies, 21*, 13-37.

Encrevé, P. (1977). Présentation : linguistique et sociolinguistique. *Langue Française, 34,* 3-16.

Falbo, C. (2005). La transcription : une tâche paradoxale. *The Interpreters' Newsletter, 13.* 25-38.

Habert, B. (2005). Portrait de linguiste(s) à l'instrument. *Texto !* [en ligne], vol. X, n°4.

Koch, P. & Oesterreicher, W. (2001). Langage oral et langage écrit. *Lexicon der romanistischen Linguistik, 1-2.* Tübingen : Max Niemeyer Verlag, 584-627.

Mondada, L. (2000). Les effets théoriques des pratiques de transcription, *LINX, 42,* 131-146.

Ochs, E. (1979). Transcription as theory, in *Developmental pragmatics,* ed. by E. Ochs & B. Schieffelin. New york : Academic Press, 43-72.

Corpus eslo : http://eslo.huma-num.fr/

Session 2.B.
Annotations, automatisation

Repérage semi-automatique du discours direct : acquisition et évaluation sur corpus Sermo (XVI-XVIIIe siècles)

Ljiljana Dolamic et Magdalena Augustyn
Université de Nauchâtel, Institut de langue et civilisation françaises
prenom.nom@unine.ch

1. Introduction

En tant que structure discursive susceptible de changements formels et structurels à travers le temps, le discours direct offre un champ privilégié pour l'étude de différents phénomènes linguistiques comme l'oral représenté, par son relatif mimétisme (Marchello-Nizia 2012, Guillot et al. 2014, 2015 pour l'ancien et le moyen français), la configuration des genres ou la réflexion plus générale autour de la relation entre le corpus et les formes du discours rapporté (Rosier 2005, López Muñoz et al. 2006) ou la ponctuation et son rôle dans l'organisation du texte et la production de sens (Boré 2009).

Nous souhaitons contribuer à cette réflexion en proposant une méthodologie de balisage semi-automatique du discours direct dans un corpus, ainsi que les premiers résultats de l'analyse pour le français préclassique et classique, un état de langue moins étudié. Nous nous appuyons sur un corpus de sermons présentant, à nos yeux, un intérêt particulier pour l'étude de la question du discours rapporté. Le genre de sermon n'a été que très peu exploité jusqu'à présent (Skupien Dekens 2014, à paraître) ; il constitue cependant une source unique pour l'histoire de la langue française, en particulier dans l'histoire des genres paralittéraires et dans son rapport avec l'oral, puisque les sermons y sont intimement liés, à des degrés divers. Le corpus SERMO, étiqueté et lemmatisé, est constitué à ce jour de 50 sermons (467 891 tokens) transcrits avec leur ponctuation d'origine, répartis entre 1550 et 1750. Le corpus est développé dans le cadre du projet « SERMO I, méthode d'annotation et d'exploitation de corpus paralittéraires pour l'analyse en linguistique diachronique » (FNS n°105212_160030), Université de Neuchâtel.

2. Discours direct dans le corpus SERMO

Certains marqueurs typographiques dont la fonction de signalisation du discours direct se stabilise progressivement, constituent un premier ancrage pour le repérage (semi)automatique de ces unités discursives (Guillot et al. 2013, Schöch et al. 2016). Cependant, en abordant la question au travers d'un corpus large et dans une perspective diachronique, nous sommes nécessairement confrontés à une diversité de formes d'intégration des séquences du discours rapporté dans le texte, de leurs différentes formes morpho-syntaxiques, ainsi qu'aux problèmes de leur délimitation et du bruit provoqué par d'éventuelles citations. Par exemple :

> Et quãd les
> gens idiots diront,Ie ne say que c'est,q̃ les plus sauans
> & plus aigus ayent cela cõme des lettres closes
> & cachettees. (Calvin_1567)

> *Qui est-ce qui dit que cela a été fait, & l'Eternel*
> *ne l'a point commandé ?* dit Ieremie au 3.
> de ses Lamentations. *Les biens & les maux*
> *ne viennent-ils pas du mandement du Tres-*
> *haut ?* On me dira , qui est-ce qui ne sait pas
> cela , que c'est Dieu qui fait tout ? Et moi ie
> vous dirai, qui est-ce qui en est persuadé comme
> il le doit ? (Superville_1700)

3. Méthode

Notre méthode de repérage semi-automatique de différentes formes du discours direct, y compris celles non marquées typographiquement, est effectuée au niveau des phrases. Pour y parvenir nous

avons construit, dans un premier temps, une chaine de traitement qui comprend : la détection des phrases, la lemmatisation et l'étiquetage morphosyntaxique. Etant donné que les résultats de ces outils sont utilisés dans notre expérimentation en tant que caractéristiques, pour la lemmatisation et l'étiquetage morphosyntaxique, ou unités de classification dans la détection de phrases, leur qualité est décisive. La précision d'étiquetage de 94.5% et de lemmatisation de 97.2% que nous obtenons à ce jour sur l'échantillon de notre corpus se révèle prometteuse. Dans un deuxième temps, nous avons créé un ensemble d'apprentissage contenant 965 phrases provenant de 50 sermons. Dans cet échantillon les segments du discours direct ont été étiquetés manuellement en aboutissant à un ensemble test de 199 phrases contenant le discours direct et 766 phrases sans le discours direct.

3.1. Extraction des caractéristiques

Nous avons d'abord extrait les caractéristiques morpho-syntactiques suivantes :

- Token («parolles», «nostre seigneur», «diront» ...)
- Lemme («PAROLE», «NOTRE-SEIGNEUR», «DIRE» ...)
- Etiquette POS («Nc», «Np», «Vvc» ...)

Un certain nombre d'autres caractéristiques, comme la longueur des phrases ou la présence du passage en italique, a été également intégré afin d'établir une liste de caractéristiques qui sera utilisée dans le test de référence. Il est important de souligner que l'usage moderne des marqueurs typographiques de discours direct comme guillemets ou tirets, n'est pas d'actualité à cette période et ces marqueurs n'apparaissent pas dans notre corpus. La présence d'italique dans la phrase a été retenue en tant que caractéristique de test de référence.

3.2. Sélection des caractéristiques

A partir de 9577 caractéristiques extraites, nous avons procédé à une sélection automatique en testant différents algorithmes (disponibles sur http://www.cs.waikato.ac.nz/ml/weka). Le Tableau 1 présente les résultats de la classification avec les méthodes de sélection les plus performantes (CfsSubsetEval et OneR). La ligne « Total » montre les résultats de la classification avec un ensemble complet des caractéristiques (9577). La classification est effectuée à l'aide de l'algorithme "Naïve Bayes" intégré dans le même logiciel. En utilisant la méthode la plus performante "CfsSubsetEval", nous avons retenu 24 caractéristiques.

3.3. Validation

Nous avons procédé à une validation de notre méthode en utilisant la validation croisée sur 10 pli à l'aide d'algorithmes d'apprentissage automatique appartenant à différentes familles : Naïve Bayes (NB), Baysian Logistic Regresion (BLR), Maximal Entropy (MaxEnt), LibSVM, Sequentil Minimal Optimization (SMO), JRip, Random Forest (RF). Le Tableau 2 présente le résultat obtenu avec ces algorithmes en les comparant au test de référence (Réf.), basé sur la présence du passage en italique. Les meilleures performances sont marquées en gras.

4. Évaluation

Afin d'évaluer la méthode de repérage du discours direct ainsi élaborée, nous avons effectué une classification d'un ensemble test inédit contenant 134 phrases (104 phrases sans le discours direct et 30 avec le discours direct) qui ne faisaient pas partie de l'ensemble d'apprentissage. Les résultats de cette évaluation (Tableau 3) montrent une légère baisse en précision par rapport aux résultats obtenus dans la validation croisée. La vérification manuelle montre, entre autres, que le système accorde trop d'importance au verbe introducteur « dire » comme caractéristique, tandis que le repérage du discours direct avec d'autres types de segments introducteurs (ex. : *il s'escria*) est moins précis. La détection des segments introducteurs et leur intégration en tant que caractéristique distincte pourrait être une solution à ce problème. Néanmoins, les résultats obtenus restent satisfaisants, la méthode présentée

TABLE 1 – **Sélection des caractéristiques**

Naïve Bayes	**précision**	**rappel**	**F-mesure**
Total	0.4850	0.4720	0.4780
CfsSubsetEval	**0.7160**	**0.6080**	**0.6580**
OneR	0.6650	0.5180	0.5820

est une base solide de développement d'un système de repérage du discours direct adapté au corpus SERMO. Nous présenterons également un premier aperçu de ce phénomène micro-structurel dans notre corpus.

Références bibliographiques

Boré, C. (2009). Remarques sur la ponctuation du discours direct dans les Contes de Perrault et de Mme d'Aulnoy. *Linx*, 60, 47-66.

Guillot, C., Lavrentiev, A., Pincemin, B., Heiden, S. (2013). Le discours direct au Moyen Âge : vers une définition et une méthodologie d'analyse. In : D. Lagorgette et P. Larrivée, *Représentations du sens linguistique 5*, Université de Savoie, 17-41, Langages, 14.

Guillot, C., Prevost, S., Lavrentiev, A. (2014). Oral représenté et diachronie : étude des incises en français médiéval. In : F. Neveu, P. Blumenthal, L. Hriba, A. Gerstenberg, J. Meinschaefer, S. Prévost, *4e Congrès Mondial de Linguistique Française, Jul 2014, Berlin, Allemagne. 8, SHS Web of Conferences.*<10.1051/shsconf/20140801284>.<halshs-01319011>

Guillot, C., Heiden, S., Lavrentiev, A., Pincemin, B. (2015). L'oral représenté dans un corpus de français médiéval (9e-15e) : approche contrastive et outillée de la variation diasystémique. In : K. Jeppesen Kragh et L. Lindschouw (éd.), *Les Variations diasystémiques et leurs interdépendances dans les langues romanes.* Actes du Colloque DIA II à Copenhague (19-21 nov. 2012). Strasbourg : Éditions de linguistique et philologie, 15-27.

López Muñoz, J. M., Marnette, S., Rosier, L. (2006). Les rôles du Discours Rapporté dans la configuration des genres. In : S. Marnette (coord.), *Dans la jungle des discours Genres de discours et Discours Rapporté.* Cadiz : Universidad de Cádiz, 18-26.

Marchello-Nizia, Ch. (2012). L'oral représenté : un accès construit à une face cachée des langues 'mortes'. In : C. Guillot, B. Combettes, A. Lavrentiev, E. Oppermann-Marsaux et S. Prévost (éd.). *Le changement en français. Etudes de linguistique diachronique.* Bern/Berlin/Bruxelles : Peter Lang, 247-264.

Rosier, L. (2005). L'analyse de discours et ses corpus à travers le prisme du discours rapporté. *Marges Linguistiques, 9*, 154-164.

Schöch, C., Schlör, D., Popp, S., Brunner, A., Henny, U., Calvo Tello, J. (2016). Straight Talk ! Automatic Recognition of Direct Speech in Nineteenth-Century French Novels. In : *Digital Humanities 2016 : Conference Abstracts,* Jagiellonian University & Pedagogical University, Kraków, 346-353.

Skupien Dekens, C. (2014). Reste-t-il des marques de l'oral dans les sermons de Calvin ? In : *"Toujours langue varie. . . " Mélanges Andres Kristol,* F. Diémoz et D. Aquino-Weber (éd.), avec la coll. de L. Grüner et A. Reusser-Elzingre. Genève : Droz, 83-97.

Skupien-Dekens, C. (à paraître). Un genre sous-exploité en histoire du français pré-classique et classique : le sermon. In : W. Ayres-Bennett, A. Carlier, J. Glikman, T. Rinsford, G. Giouffi & C. Skupien-Dekens, (eds.), *Nouvelles voies d'accès au changement linguistique.* Paris : Garnier.

TABLE 2 – **Résultats de la validation**

	Précision			**Rappel**			**F-mesure**		
Algorithme	**DD**	**noDD**	**mp**	**DD**	**noDD**	**mp**	**DD**	**noDD**	**mp**
Réf.	0.5128	0.8714	0.7975	0.5025	0.8760	0.7989	0.5076	0.8736	0.7982
NB	0.7160	0.9020	0.8640	0.6080	0.9370	0.8690	0.6580	0.9190	0.8650
BLR	0.8480	0.8960	0.8860	0.5630	0.9740	0.8890	0.6770	0.9330	0.8800
MaxEnt	0.8600	**0.9080**	**0.8980**	**0.6180**	0.9740	**0.8980**	**0.7190**	**0.9400**	**0.8940**
LibSVM	0.8580	0.8860	0.8630	0.5180	0.9780	0.8830	0.6460	0.9300	0.8710
SMO	**0.8790**	0.8780	0.8780	0.4720	**0.9830**	0.8780	0.6140	0.9270	0.8630
JRip	0.7930	0.9020	0.8790	0.5980	0.9600	0.8850	0.6820	0.9300	0.8790
RF	0.8010	0.9040	0.8830	0.6080	0.9610	0.8830	0.6910	0.9320	0.8820

TABLE 3 – **Classification de l'ensemble du test inédit**

	Précision			**Rappel**			**F-mesure**		
Algorithme	**DD**	**noDD**	**mp**	**DD**	**noDD**	**mp**	**DD**	**noDD**	**mp**
Réf.	**0.6800**	0.8700	0.8200	0.5000	**0.9300**	0.8400	0.5800	**0.9000**	0.8300
NB	0.6220	**0.9280**	**0.8590**	**0.7670**	0.8650	**0.8430**	**0.6870**	***0.8960***	**0.8490**
BLR	0.6210	0.8860	0.8260	0.6000	0.8940	0.8280	0.6100	0.8900	0.8270
MaxEnt	0.6000	0.8850	0.8210	0.6000	0.8850	0.8210	0.6000	0.8850	0.8210
LibSVM	***0.6670***	0.8880	0.8380	0.6000	0.9130	0.8430	0.6320	0.9000	0.8400
SMO	***0.6670***	0.8730	0.8270	0.5330	***0.9230***	0.8360	0.5930	0.8970	0.8360
JRip	0.6300	0.8790	0.8270	0.5670	0.9040	0.8280	0.5960	0.8910	0.8250
RF	0.5860	0.8760	0.8110	0.5670	0.8850	0.8130	0.5760	0.8800	0.8120

Du petit fait à la *doxa* : annotation automatique des anecdotes dans le discours critique sur Molière

Elodie Bénard et Motasem Alrahabi
Université Paris-Sorbonne - OBVIL
Université Paris-Sorbonne Abu Dhabi
{elodie.benard@paris-sorbonne.fr, motasem.alrahabi@gmail.com

1. Anecdotes : rôle, définition et corpus

Nos propres jugements de Molière et nos lectures de son théâtre sont tributaires de la tradition biographique et critique qui s'est constituée après sa mort et qui est caractérisée par la présence massive d'anecdotes. Celles-ci ne sont pas présentes uniquement dans les recueils spécialisés et dans les biographies qui en font leur miel, mais elles se retrouvent dans tous les récits, commentaires et jugements, tant elles sont constitutives du discours sur Molière. Les anecdotes sont constamment répétées, même pour être réfutées. Elles informent en somme toute la réception de l'auteur comique. En français contemporain, l'anecdote est définie comme le « récit bref d'un petit fait curieux » [Rey, 2005]. Cette acception soulève deux difficultés : le seuil de narrativité et la nature du fait raconté — l'adjectif « curieux » pouvant recouvrir aussi bien les sens « piquant, amusant » que « rare, secret », « intéressant, révélateur ». Une définition fondée sur le seuil de narrativité [1] [K. Abiven, 2015] amène à exclure des séquences peu narrativisées, qui font pourtant allusion à un micro-récit originel, identifiable, et qui attestent son empreinte dans le discours critique. En ce qui concerne la nature du fait rapporté, deux conditions sont nécessaires : le fait ne peut pas être authentifié de façon certaine et il contribue à la construction de l'« écrivain imaginaire », c'est-à-dire, l'écrivain tel qu'il se fait représenter [Diaz, 2007]. Dans notre approche, nous appelons anecdote toute séquence qui a la forme d'un micro-récit ou renvoie sur le mode allusif à un micro-récit, dont la véracité est sujette à caution et qui est révélateur d'une facette de l'instance imaginaire construite par les biographes et les critiques. Cette définition, relativement extensive, n'est pas fondée sur le degré de narrativité et inclut aussi des séquences de longueur variable — les plus courtes contiennent une seule phrase et les plus longues peuvent être composées de plusieurs paragraphes. Elle repose sur trois déterminations essentielles de l'anecdote : sa mise en circulation, sa véridicité problématique et sa fonction symbolique. Ces trois caractéristiques peuvent avoir une traduction formelle, en particulier par des marques linguistiques de la citation de discours ou de la modalisation. Cependant, ces marques peuvent être gommées : « L'une des fonctions les plus fortes de l'idéologie est justement de nous faire croire que tout usage relève d'un usage personnel, d'une appropriation et non d'une réappropriation. La circulation visera donc autant à l'effacement des strates énonciatives qu'à son marquage selon les conditions de circulation du discours. » [Rosier, 2003]. En cela, l'anecdote est proche de discours comme la rumeur. Voici un exemple d'anecdote sur Molière :

> *On rapporte que Molière, en faisant répéter cette pièce, parut mécontent des acteurs qui y jouaient, et principalement de Mlle Beauval, qui représentait le personnage de Toinette.*
> *(Corpus Molière – OBVIL)*

Notre corpus est composé de textes écrits après la mort de Molière jusqu'aux premières décennies du XXe siècle : des recueils d'anecdotes du XVIIIe siècle, la critique moliéresque produite aux XVIIIe et XXe siècles (biographies, ouvrages sur différents aspects de la vie de Molière et de son théâtre et chapitres d'histoires de la littérature) et les paratextes des éditions des œuvres de Molière. Ces textes, actuellement au nombre de 104, ont été préalablement numérisés dans le cadre du Labex OBVIL de Paris-Sorbonne [Alexandre, 2016] et sont librement accessibles [2]. Eu égard à la taille du corpus (2,5 millions de mots) et à notre objet qui possède des caractéristiques linguistiques identifiables [K. Abiven, 2015], il est pertinent de recourir à l'annotation automatique des textes. Celle-ci permet dans notre contexte de viser un repérage qui soit le plus complet possible, indispensable pour montrer la

1. Seuil en deçà duquel on n'a pas affaire à un récit minimal, défini comme la relation d'actions temporellement ordonnées (t→t+n) et tendue entre un début et une fin permettant de nouer et de dénouer une action [J. Bres, 1988].
2. http://obvil.paris-sorbonne.fr/corpus/moliere/critique/

place occupée par les anecdotes dans le discours critique, mettre en évidence l'invention continuelle de petits récits (favorisée par le caractère répétitif et paradigmatique des schèmes anecdotiques) et repérer les variations formelles et thématiques d'une anecdote au fil du temps.

2. Marqueurs linguistiques des anecdotes

Nous avons isolé un échantillon de textes ($\approx$ 7% de la taille totale du corpus) pour analyser les anecdotes et en identifier les marqueurs. Nous avons alors observé que les anecdotes, en tant que forme de circulation du discours, sont souvent introduites par des indicateurs relevant de la catégorie du médiatif [Guentcheva, 1996], qui permettent à l'énonciateur de se figurer comme passeur d'une information et parfois comme étape dans la chaîne des discours rapportés. Exemples : *au dire de, à en croire, sur la foi de, selon le témoignage de, entendre dire, ouï dire, rapporter, répéter, révéler,* etc.

> *Mais ayant ouï dire que Molière voulait faire une comédie des Femmes savantes, elle supprima sa dissertation. (Corpus Molière – OBVIL)*

Ces indicateurs sont pertinents, mais ne suffisent pas toujours pour identifier les anecdotes. Nous avons donc identifié des indices complémentaires dans le contexte afin de lever l'ambiguïté sur certains indicateurs ou bien d'affiner la catégorisation. Différentes classes d'indices ont ainsi été créées comme les noms propres (*Grimarest, Tallemant...*) et les indices d'encadrement du micro-récit (*un jour, un autre soir...*). Si les signaux de l'ouverture du micro-récit comme « un jour » ou « un soir » sont généralement placés au tout début de la première phrase ou en zone préverbale, la place des marqueurs relevant du médiatif varie. Les premiers tests ont soulevé une autre difficulté : environ une anecdote sur dix dans le corpus de test ne concerne pas Molière :

> *Racine lui avait confié le rôle d'Oreste ; et ce rôle, suivant une tradition populaire, fut la cause de sa mort ; il [*i. e. *l'acteur Montfleury] se rompit, dit-on, une veine, par les efforts prodigieux qu'il fit pour bien rendre la scène des fureurs. (Corpus Molière – OBVIL)*

Racine lui avait confié le rôle d'Oreste ; et ce rôle, suivant une tradition populaire, fut la cause de sa mort ; il [i. e. l'acteur Montfleury] se rompit, dit-on, une veine, par les efforts prodigieux qu'il fit pour bien rendre la scène des fureurs. (Corpus Molière – OBVIL) Nous avons donc ajouté à la liste des indices complémentaires les variantes du nom « Molière » : *jeune tapissier, poète comédien, patron de la troupe, Poquelin,* etc. Cette première approche va nous aider à mieux cerner les anecdotes sur Molière, mais ne va pas résoudre définitivement le problème, car environ 55% des anecdotes sur Molière dans le corpus analysé ne comportent pas dans leurs phrases introductrices de mention de son nom ou de l'une de ses variantes. Au total, nous avons collecté pour cette tâche autour de 60 marqueurs linguistiques.

3. Annotation automatique

Pour l'annotation automatique, nous avons utilisé excom2, un outil d'annotation à base de règles et de marqueurs linguistiques de surface [Alrahabi, 2010]. Ceux-ci sont organisés dans des catégories sémantiques et/ou discursives : *opinions, définitions, comparaisons, conclusions, hypothèses, expressions ironiques,* etc. La présence dans une phrase d'un indicateur déclenche les règles associées qui explorent le contexte [3] et vérifient la présence ou l'absence d'indices complémentaires [Desclés, 2006]. Voici un exemple de règle :

3. Pour la création des espaces de recherche dans un texte, excom2 effectue un prétraitement consistant à segmenter les documents en sections, paragraphes et phrases [Alrahabi, 2010].

SI dans une phrase on a un indicateur comme : *en croit, à en croire, s'il faut l'en croire...*

ET SI dans le contexte *Après*, on a un indice comme *un nom propre (Grimarest...), une fonction (auteur...), un terme comme Certains...*

ET SI dans les contextes *Avant* ou *Après* nous n'avons pas de négation comme *ne, pas, rien, plus...*

ALORS annoter la phrase en cours.

EXEMPLE *A en croire Grimarest, l'original de ce maître de philosophie serait Rohault, un des plus zélés et des plus célèbres disciples de Descartes, et en même temps ami de Molière.*

Les règles dans excom2 peuvent être organisées selon un ordre de priorité et utiliser les résultats d'autres règles. Pour les Anecdotes, nous avons créé 5 règles que nous avons associées aux différents marqueurs linguistiques (indicateurs et indices). Une première phase de test sur l'échantillon du corpus était nécessaire pour la stabilisation des règles.

4. Évaluation des résultats

Afin d'évaluer la qualité des annotations automatiques, nous nous sommes focalisés dans un premier temps sur le calcul de la *précision*. Dans cette perspective, nous avons annoté avec excom2 le reste du corpus ($\approx$ 93%) et obtenu 1096 annotations. Ensuite, nous avons demandé à une personne experte d'évaluer les sorties selon un guide d'annotation. Pour chaque phrase annotée, l'évaluatrice devait choisir entre trois étiquettes que nous présentons dans le tableau suivant avec les résultats de l'évaluation :

Etiquette	**Nombre d'annotations correctes**	**Précision**
Circulation du discours	963	87.9%
Anecdote	592	54.1%
Anecdote qui porte sur Molière	425	38.8%

Vu la particularité du phénomène langagier des anecdotes et la simplicité de notre approche par analyse de surface, nous considérons que ces premiers résultats sont très encourageants et méritent d'être améliorés.

5. Discussion et perspectives

Dans la perspective de l'étude des jugements de valeur sur Molière, les anecdotes sont une voie d'entrée féconde dans le corpus critique car elles éclairent la manière dont s'est élaborée et transformée la *doxa* sur Molière. L'approche que nous avons adoptée nous a permis, avant toute autre chose, de découvrir de nouvelles anecdotes non encore étudiées. Elle nous a fourni une matière abondante et des données quantitatives pour mieux cerner l'objet d'étude.

Nous constatons notamment que les marqueurs linguistiques du récit (comme *un jour*) sont présents dans 20% environ des séquences identifiées comme anecdotes, le reste est repéré grâce à des marqueurs qui renvoient à la circulation du discours et/ou expriment le doute sur sa véridicité. Déterminer précisément, c'est-à-dire quantitativement, le type de marqueurs qui introduit l'anecdote permet de mieux appréhender sa fonction dans le discours biographique et critique en tant qu'élément d'une construction argumentative visant à donner une certaine image de l'auteur et de son œuvre. De plus, les résultats permettent de confronter les séquences que nous identifions comme anecdotes et les séquences qui présentent des traits linguistiques semblables, mais ne relèvent pas de la définition retenue : là encore, le rapprochement de l'anecdote avec des catégories comme le jugement ou l'opinion aide à mieux cerner la spécificité de l'objet.

Afin d'améliorer les résultats, nous allons élargir la couverture des ressources linguistiques, en testant de nouveaux marqueurs comme *dire, déclarer, informer,* etc. ou bien des adjectifs comme *extravagant, inattendu, curieux,* etc. Concernant les anecdotes qui ne sont pas en rapport avec Molière, nous avons constaté que les indices complémentaires que nous avons ajoutés (*Poquelin...*) ne sont pas suffisants : sur la totalité des anecdotes repérées, le système arrive à identifier uniquement 72%

d'anecdotes qui concernent Molière. Nous envisageons donc d'explorer d'autres pistes qui vont nous confronter à la problématique du repérage des entités nommées et de l'anaphore.

Les marqueurs des anecdotes dans notre approche permettent de localiser uniquement la phrase qui introduit ou qui conclut l'anecdote, avec, selon le cas, le contenu de l'anecdote. Or, nous avons constaté dans les résultats que la longueur d'anecdote est de 5 à 6 phrases en moyenne. Nous allons donc rechercher des solutions pour délimiter les frontières des anecdotes. Enfin, notre objectif à court terme est de créer et de diffuser un corpus de référence (*Gold Standard Corpus*) où les anecdotes sont manuellement vérifiées et annotées. Ceci nous permettra, entre autres, de mesurer le *rappel* et d'améliorer l'annotation automatique.

Références bibliographiques

Abiven, K. (2015). *L'Anecdote ou la fabrique du petit fait vrai. De Tallemant des Réaux à Voltaire* (1650-1750), Paris, Classiques Garnier, 2015.

Alexandre, D. (2010). Études littéraires et calcul numérique. Présentation, *Revue d'histoire littéraire de la France,* 3, juillet-septembre 2016, Paris. p. 517-520.

Alrahabi, M. (2010). *EXCOM-2 : plateforme d'annotation automatique de catégories sémantiques. Applications à la catégorisation des citations en français et en arabe.* Thèse de doctorat, sous la direction du Prof. Jean-Pierre Desclés, Université Paris-Sorbonne.

Bres, J. (1988) À la recherche de la narrativité : éléments pour une théorisation praxématique, dans J. Bres (dir.), *Du récit, encore, Cahiers de praxématique,* Montpellier, Université Paul Valéry-Montpellier III, 11, p. 75-100.

Desclés, J.-P. (2006). *Contextual Exploration Processing for Discourse Automatic Annotations of Texts,* Actes de FLAIRS 2006, Florida, USA

Diaz, J. L. (2007). *L'Écrivain imaginaire. scénographies auctoriales à l'époque romantique,* Paris, H. Champion.

Guentcheva, Z. (1996). *L'Énonciation médiatisée,* Louvain, Peeters, Paris

Rey, A. (2005). *Dictionnaire culturel en langue française,* Paris.

Rosier, L. (2003). *Du discours rapporté à la circulation des discours : l'exemple des dictionnaires de "critique ironique",* dans Lopez-Muñoz, J.-M., Marnette, S. et L. Rosier (éds). « Formes et stratégies du discours rapporté : Approches linguistique et littéraire des genres de discours », Estudios Lengua y Literatura francesas, 14, p. 63-82.

Annotation manuelle d'expressions référentielles : expérimentations pour simplifier les prises de décisions et optimiser le processus

Frédéric Landragin , Juliette Potier et Meryl Bothua
Laboratoire Lattice
CNRS, ENS, Université de Paris 3, Université Sorbonne Paris Cité, PSL Research University
frederic.landragin@ens.fr

1. Introduction

La constitution et l'annotation manuelle d'un corpus regroupant des textes écrits de différents genres textuels et de différentes époques fait partie des objectifs du projet ANR Democrat [1]. L'objet d'étude est double : les expressions référentielles et les chaînes de référence. La taille envisagée est d'environ un million de mots, soit de l'ordre de 200 000 expressions référentielles annotées, ce qui représente une tâche d'annotation importante. Les arguments sont les suivants : il s'agit premièrement de fournir un corpus de référence qui serve à toute la communauté, deuxièmement de permettre des analyses statistiques (textométriques) avec des données en quantité suffisante, et troisièmement de nourrir un ou plusieurs systèmes d'apprentissage artificiel, de manière à ouvrir la voie de la détection automatique des expressions référentielles et des chaînes de référence en français, suite aux expérimentations déjà réalisées avec le corpus Ancor (Désoyer *et al.*, 2014), seul corpus de taille comparable (115 000 anaphores annotées).

Compte tenu de la taille de corpus envisagé, la procédure d'annotation manuelle doit être la plus efficace possible, afin de ne pas faire appel à des moyens humains démesurés. Nous décrivons dans cette présentation les discussions et les expérimentations qui ont permis de définir une procédure d'annotation rationnelle pour l'annotation des mentions, c'est-à-dire pour la phase la plus coûteuse de l'annotation – et celle sur laquelle reposeront les phases suivantes, comme celle d'annotation des chaînes de référence. Nos choix reposent sur l'expérience des annotateurs du corpus Ancor (Muzerelle *et al.*, 2014) et du corpus MC4 (Landragin, 2011), sur des calculs d'accords inter-annotateurs et sur des expérimentations chronométrées comparatives.

2. La référence et les expressions référentielles

2.1. Annoter quoi et comment ?

La référence est un objet linguistique très vaste, qui a fait l'objet de très nombreuses publications (Charolles 2002). Le projet MC4 (Landragin, 2011) s'était intéressé aux multiples facteurs morphologiques, syntaxiques, sémantiques et pragmatiques qui interviennent lors de la résolution des références, c'est-à-dire l'attribution d'un référent à une expression référentielle, en incluant l'attribution d'un antécédent à une anaphore. La procédure d'annotation résultante avait impliqué l'annotation manuelle de ces facteurs, du moins d'une sélection d'une dizaine de facteurs considérés comme déterminants. Au final, le corpus n'a pas dépassé 5 000 expressions référentielles annotées. Pour le projet Democrat, il n'était pas question de reproduire une procédure aussi détaillée, et ce d'autant plus que l'annotation de certains de ces facteurs est automatisable. Nous avons ainsi choisi d'annoter uniquement le résultat de la résolution de la référence. Deux possibilités apparaissent ici : soit on saisit, pour chaque expression, un identifiant du référent ; soit on regroupe les expressions en chaînes (anaphoriques et/ou coréférentielles), ce qui fait l'économie des identifiants des référents mais nécessite de construire des chaînes, autrement dit des objets non liés à un et un seul marquable.

Une première expérimentation a permis de comparer les deux méthodes et a soulevé l'importance décisive de l'ergonomie de l'outil d'annotation utilisé : comme il est possible de déduire automatiquement les chaînes à partir d'une annotation des mentions en identifiants, seule compte la rapidité d'action. Or, quand l'outil est bien choisi et permet la complétion automatique de l'identifiant en cours de saisie, il s'avère que la méthode à base d'identifiants est plus rapide que celle à base de construction de chaînes. En effet, manipuler un objet couvrant potentiellement le texte entier est bien plus

1. Democrat, « description et modélisation des chaînes de référence : outils pour l'annotation de corpus (en diachronie et en langues comparées) et le traitement automatique », projet ANR-15-CE38-0008.

délicat et propice à des erreurs que saisir des identifiants localement, au niveau du marquable qu'est l'expression référentielle. Plusieurs outils ont été testés (Mmax2, Glozz, Analec) et c'est finalement l'implémentation de la complétion dans l'outil Analec qui a permis la plus grande efficacité (Landragin *et al.*, 2012).

2.2. L'accord inter-annotateurs pour la référence

S'il n'est pas possible de valider une procédure d'annotation (à moins de disposer d'un *gold standard*), il est possible d'évaluer sa reproductibilité, ce qui donne une indication précieuse sur l'intérêt des annotations d'un corpus (Mathet et Widlöcher, 2016). Ceci se fait en impliquant plusieurs annotateurs – 2 à 3 dans nos expérimentations – puis en calculant l'accord inter-annotateurs. Plusieurs indicateurs statistiques sont couramment utilisés : α, π, κ et désormais γ (Mathet *et al.*, 2015). Même si nous avons calculé lors de chaque comparaison importante de procédures l'ensemble des indicateurs, nous soulignons qu'aucune expérimentation n'a permis d'obtenir des scores très élevés : autour de 0,53 pour γ dans les premières expérimentations, autour de 0,73 dans les dernières (α entre 0,662 et 0,733, π entre 0,66 et 0,732, κ entre 0,661 et 0,733). Ceci est dû à la nature de l'objet d'étude : la coréférence et l'anaphore sont des phénomènes complexes, pour lesquels les interprétations peuvent varier d'un annotateur à l'autre sans pour autant que les annotations en deviennent inutisables. Les corpus constitués dans d'autres langues que le français font face à des scores similaires aux nôtres, et la communauté s'est habituée à des taux d'accord modestes (Artstein et Poesio, 2008).

3. Identifier des stratégies d'annotation

Nos premières expérimentations d'annotation se sont déroulées sur des textes narratifs, en l'occurrence des extraits de romans libres et gratuits, disponibles sur la plateforme *wikisource*. Il s'agit donc de textes littéraires, et les discussions ont vite porté sur le filtrage ou non des référents : faut-il annoter toutes les expressions référentielles, ou seulement celles qui réfèrent à des personnages humains, à des êtres animés, à des objets concrets, etc. Nous constatons par exemple que les expressions référentielles temporelles sont très peu reprises et ne forment donc que rarement des chaînes de référence intéressantes à étudier. Dans ce cas, faire l'impasse sur l'annotation de ces « singletons » permettrait d'aller plus vite à l'essentiel. Sauf que : 1. l'annotation systématique de toutes les expressions référentielles permet de nourrir un système d'apprentissage dédié à la détection des expressions référentielles (ce qui, en TAL, est une tâche très complexe, différente de celles consistant à détecter les entités nommées et les pronoms anaphoriques) ; 2. il n'est pas possible de savoir si l'expression en cours d'annotation va être reprise ultérieurement ou non (autrement dit la tâche peut comporter des retours en arrière dans le texte si l'annotateur s'aperçoit qu'il a oublié une expression – initialement considérée comme singleton) ; 3. se demander à chaque expression si elle a des chances d'être un singleton ou de faire partie d'une chaîne de référence va à l'encontre de l'aspect « robotique » et efficace de l'annotation : se poser trop de questions est parfois contre-productif, et il vaut mieux tout annoter en se posant moins de questions. Nous avons convenu de rendre le choix des identifiants de référents le plus rapide possible pour les expressions peu susceptibles d'être reprises. Pour les expressions qui resteront clairement des singletons, nous employons un code dédié (« SI » comme singleton), ce qui permet d'augmenter encore l'efficacité.

4. Pré-annotation : aide ou gêne ?

L'aspect robotique du repérage des expressions référentielles (pas de leur attribution d'un référent) peut être encouragé en utilisant un système de TAL en tant que pré-annotateur. Mais encore faut-il trouver un trouver un système de TAL qui soit adapté au français et dont le taux d'erreur ne soit pas une entrave à l'annotation : quand les erreurs sont nombreuses, l'annotateur a vite l'impression de passer son temps à les corriger plutôt qu'à exploiter directement les pré-annotations. Or il n'existe pas de système de détection automatique des expressions référentielles, et il nous faut nous rabattre sur la détection des entités nommées (ce qui est une tâche plus réduite, cf. Nouvel *et al.*, 2015), la détection des anaphores (ce qui est également très réducteur) ou bien sûr la détection des chaînes de référence – sauf que le seul système disponible pour le français, RefGen (Todiraşcu et Longo, 2011) a

des performances moyennes. En fin de compte, le système le plus proche du résultat souhaité est tout simplement un détecteur de *chunks* nominaux. Après un comparatif des performances des différents outils disponibles, notre choix a porté sur le *chunker* nominal de SEM (Tellier *et al.*, 2012).

Le principal avantage d'un *chunker* nominal est qu'il permet à l'annotateur de n'oublier aucune expression référentielle. Mais quelques inconvénients viennent contrebalancer cet avantage : 1. tous les *chunks* nominaux d'un texte ne réfèrent pas, donc le *chunker* produit du bruit qui peut perturber l'annotateur (« il » impersonnel, mention non référentielle de partie du corps comme « avoir la grosse tête », etc.) ; 2. un *chunker*, par définition, repère des portions de texte non enchâssées – or des expressions référentielles peuvent s'enchâsser, comme les compléments du nom. Des adaptations des résultats du *chunker* sont donc nécessaires et, là aussi, tout repose sur l'ergonomie de l'outil d'annotation utilisé. Ainsi, un outil qui permet de rectifier facilement les frontières d'un marquable peut avantager l'exploitation d'une pré-annotation en *chunks*.

5. Conclusion

Les objectifs textométriques et TAL de notre corpus encouragent à privilégier certaines stratégies d'annotation par rapport à d'autres. L'intérêt de disposer d'une annotation – même minime – de toutes les expressions référentielles nous incite à diriger la procédure d'annotation dans ce sens. C'est une décision qui privilégie la référence à la coréférence, mais plusieurs expérimentations chronométrées (tranches de 30 minutes d'annotation avec des conditions différentes) ont montré que le surcoût en temps d'annotation reste raisonnable compte tenu du gain final : dans le pire des cas (sur 7 chronométrages), la perte de temps est de 50%. Quant à l'utilisation d'un *chunker* pour disposer d'un pré-repérage des expressions référentielles, elle a été laissée au choix de l'annotateur : comme beaucoup d'aspects de la tâche d'annotation, l'appropriation de la procédure se fait mieux si une certaine souplesse est autorisée, sous condition bien entendu que les annotations finales se soient pas impactées, ce qui est le cas ici.

Références bibliographiques

Artstein, R., Poesio, M. (2008). Inter-Coder agreement for Computational Linguistics, *Computational Linguistics, 34,* 555-596.

Carletta, J. (1996). Assessing agreement on classification tasks : the kappa statistic. *Computational Linguistics, 22,* 249-254.

Charolles, M. (2002). *La référence et les expressions référentielles en français.* Paris : Ophrys.

Désoyer, A., Landragin, F., Tellier, I., Lefeuvre, A., Antoine, J.-Y. (2014). Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus Ancor. *Traitement Automatique des Langues, 55(2),* 97-121.

Heiden, S., Magué, J.-P., Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In : *Proceedings of Tenth International Conference on the Statistical Analysis of Textual Data,* Vol. 2, 1021-1032.

Krippendorff, K. (2012). *Content analysis : an introduction to its methodology (third edition).* Thousand Oaks : Sage Publishing.

Landragin, F. (2011). Une procédure d'analyse et d'annotation des chaînes de coréférence dans des textes écrits. *Corpus, 10,* 61-80.

Landragin, F., Poibeau, T., Victorri, B. (2012). Analec : a New Tool for the Dynamic Annotation of Textual Data. In : *Proceedings of the Eighth International Conference on Language Resources and Evaluation,* Istanbul, Turquey, 357-362.

Mathet, Y., Widlöcher, A. (2016). Évaluation des annotations : ses principes et ses pièges. *Traitement Automatique des Langues, 57(2),* 73-98.

Mathet, Y., Widlöcher, A., Métivier, J.-P. (2015). The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment. *Computational Linguistics, 41(3),* 437-479.

Müller, C., Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In : Braun, S., Kohn, K., Mukherjee, J. (Eds.). *Corpus technology and language pedagogy : New resources, new tools, new methods.* Frankfurt : Peter Lang.

Muzerelle, J., Lefeuvre, A., Schang, E., Antoine, J.-Y., Pelletier, A., Maurel, D., Eshkol, I., Villaneau, J. (2014). Ancor centre, a large free spoken french coreference corpus : description of the resource and reliability measures. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation,* Reykjavik, Iceland.

Nouvel, D., Ehrmann, M., Rosset, S. (2015). *Les entités nommées pour le traitement automatique des langues.* Londres : Éditions ISTE.

Tellier, I., Duchier, D., Eshkol, I., Courmet, A., Martinet, M. (2012). Apprentissage automatique d'un chunker pour le français. In *Actes de la 19e Conférence sur le Traitement Automatique des Langues Naturelles,* Grenoble.

Todiraşcu, A., Longo, L. (2011). RefGen, outil d'identification automatique des chaînes de référence en français. *18e Conférence sur le Traitement Automatique des Langues Naturelles,* session des démonstrations industrielles, Montpellier.

Widlöcher, A., Mathet, Y. (2009). La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. In : *Actes de la 16e Conférence sur le Traitement Automatique des Langues Naturelles,* Senlis.

Session 3.A.
Constitution de ressources

Vers des ressources électroniques interconnectées : Lexica, les dictionnaires de la collection Pangloss

Rémy Bonnet [1], Céline Buret [2], Alexandre François [2], Benjamin Galliot [2], Séverine Guillaume [2], Guillaume Jacques [1], Aimée Lahaussois [3], Boyd Michailovsky [2] et Alexis Michaud [2]
[1]Centre de Recherches Linguistiques sur l'Asie Orientale (CRLAO, CNRS / EHESS / INALCO)
[2]Langues et civilisations à tradition orale (Lacito, CNRS / Université Sorbonne Nouvelle / INALCO)
[3]Histoire des Théories Linguistiques (HTL, CNRS / Université Paris Diderot / Université Sorbonne Nouvelle)
{bonnet.remy et buret.celine et b.g01lyon et rgyalrongskad et aimeelah et boyd.michailovsky} @gmail.com
{alexandre.francois et severine.guillaume et alexis.michaud}@cnrs.fr

1. Introduction : vers des ressources électroniques interconnectées

1.1. Données et corpus dans l'exploration de la diversité linguistique

Traditionnellement, la linguistique de terrain vise à la production de grammaires, dictionnaires, et recueils de textes. Ces trois éléments forment ce que l'on appelle la « trilogie boasienne » (Foley, 1999) par référence au travail fondateur de Franz Boas (Boas, 1902 ; Boas & Swanton, 1911). Rien de computationnel dans cette méthode, formulée à une époque où les chercheurs publiaient leurs travaux sous forme imprimée. Mais un siècle plus tard, les technologies numériques permettent une avancée décisive : par l'ajout de la composante multimédia (enregistrements audio et vidéo), la trilogie est devenue *tétralogie* (Musgrave & Thieberger, 2014). Ce tournant a été pris au Lacito dès 1994, par la fondation de la collection Pangloss (Jacobson, Michailovsky, & Lowe, 2001 ; Michailovsky et al., 2014) – archive multimédia en ligne actuellement en pleine expansion.

L'usage des nouvelles technologies va bien plus loin que la simple publication en ligne de travaux autrefois imprimés. Ce qui est désormais crucial, c'est l'établissement de liens dynamiques entre les quatre volets de la tétralogie : demain, dictionnaires et grammaires pourront non seulement être inter-connectés, mais aussi liés aux textes qui forment le cœur des données linguistiques, ainsi qu'aux enregistrements audio et vidéo de parole spontanée. Plus que de fixer une langue au moyen de l'imprimé, il s'agit désormais de l'offrir à des modes nouveaux de navigation, en exploitant tout le potentiel de corpus en ligne, y compris par des traitements statistiques. Le projet de *tétralogies connectées* et de grammaires électroniques a été formulé clairement (Maxwell, 2012 ; Nordhoff, 2008). La présente communication expose l'état d'avancement de réalisation de dictionnaires en ligne, étape dans l'entreprise qui consiste à porter le projet de *tétralogies connectées* au stade des réalisations pratiques.

1.2. La collection Pangloss et la « linguistique assistée par ordinateur »

La collection Pangloss regroupe un ensemble de ressources multimédia. Enregistrements audio ou vidéo effectués sur le terrain par les chercheurs qui étudient des langues dites « rares », transcriptions annotées des contenus des enregistrements, dictionnaires multimédias et, en projet, des grammaires. Plus de 130 langues, des centaines d'heures d'enregistrements, un millier de documents annotés, le tout librement consultable.

Notre travail s'inscrit dans le contexte d'une linguistique assistée par ordinateur. « Les humanités numériques vont dans le sens d'une approche dans laquelle une distinction nette entre humanités d'une part et informatique de l'autre n'est plus constructive » (Collins et al., 2015, p. 10). Dans ce contexte, chaque ressource peut être directement connectée aux autres ressources. Ainsi, de nombreux enregistrements audio sont synchronisés avec les textes annotés. Les dictionnaires eux même sont reliés à des enregistrements audio de la collection et les exemples textuels des entrées des dictionnaires ont vocation à être accompagnés de la possibilité d'écouter l'exemple en contexte.

À l'ère du numérique, l'interconnexion des ressources ainsi que les outils d'exploitation automatique facilitent grandement la confrontation des hypothèses avec les données, ce qui encourage la réalisation d'implémentations logicielles des modèles linguistiques.

2. Lexica : présentation des dictionnaires en ligne

Le nom « Lexica » a été adopté pour la dimension lexicographique de la collection Pangloss.

2.1. Pourquoi de nouveaux outils ? La bibliothèque PYLMFLIB

Pour la création de dictionnaires, les outils tels que Toolbox, Fieldworks et LexiquePro, développés par le *Summer Institute of Linguistics*, sont largement utilisés par les linguistes « de terrain », mais ils manquent de flexibilité, en particulier pour l'inclusion des paradigmes de conjugaison des verbes et la conversion automatique d'une orthographe en une autre. Le format utilisé pour les dictionnaires Toolbox (MDF) présente en outre le désavantage d'avoir une structure implicite et ambiguë.

Un travail de développement informatique a donc été réalisé. Une librairie en python, PYLMFLIB (https://pypi.python.org/pypi/pylmflib/1.0), a été développée par l'un des auteurs de la présente communication (Céline Buret). Cette librairie implémente en XML la norme lexicographique LMF (Francopoulo, 2013), conçue comme un format pivot (Romary, 2013), avec des outils de conversion du format MDF vers LaTeX, HTML et DOC. Des versions PDF des dictionnaires sont générées à partir de LaTeX.

2.2. Réalisations : les dictionnaires

Trois dictionnaires multimédias (« *talking dictionaries* ») sont actuellement disponibles via l'interface de la collection Pangloss (http://lacito.vjf.cnrs.fr/pangloss/dictionaries/) : japhug, khaling, et limbu. Trois autres dictionnaires (mwotlap, na et teanu) ne comportent pas encore de liens vers des enregistrements à l'heure actuelle : pour ces dictionnaires, le travail d'interconnexion avec les ressources audio et textuelles de la collection Pangloss est en cours.

Le dictionnaire limbu (limbu-népali-anglais), conçu il y a une quinzaine d'années, a été le premier dictionnaire de la collection Pangloss (Michailovsky, 2002, voir également 2011). Il s'agit d'un dictionnaire multimédia pour lequel les exemples des entrées du dictionnaire sont directement reliés à des enregistrements de récits et d'élicitation de vocabulaire, synchronisés eux-mêmes avec leur annotation textuelle et déposés dans la collection Pangloss. Premier dictionnaire « connecté », il est en cours de conversion au format LMF.

Le dictionnaire japhug (japhug-chinois-français) comporte plus de 7 000 entrées. C'est le premier dictionnaire de cette langue, d'une grande importance pour l'étude de la famille sino-tibétaine (Jacques, 2016). Il inclut plus de 4 000 fichiers audio d'exemple de phrases, qui sont intégrées au fichier PDF et disponibles dans la version en ligne.

Le dictionnaire khaling (khaling-népali-anglais) recense tous les verbes primaires de cette langue, avec des exemples de phrases et des définitions en népali et en anglais et des tableaux de conjugaison pour tous les verbes. (Au sujet de la morphologie de cette langue, voir : Jacques, 2015 ; Jacques, Lahaussois, Michailovsky, & Bahadur Rai, 2012.) Une version papier a été publiée localement par la communauté khaling en février 2016 à Kathmandou, en plus de la version PDF intégrant les fichiers audio et la version en ligne.

La librairie est en train d'être pourvue de nouveaux outils (conversion des dictionnaires vers le format Android, interface graphique) pour permettre la production de dictionnaires multimédias dans d'autres langues étudiées par les linguistes de nos laboratoires et d'ailleurs. Une version « LMF » des dictionnaires mwotlap et teanu, et leurs déclinaisons sous forme HTML et PDF, sont en cours de finalisation. (Au sujet de ces langues, voir François, 2003, 2009.)

3. Morphologie et morphotonologie : générateurs de paradigme et projets de modélisation

Les langues que nous étudions présentent des caractéristiques intéressantes sur le plan de la morphologie et de la morphophonologie (notamment en ce qui concerne les tons). Les règles morphologiques et morphophonologiques peuvent être implémentées, ce qui permet une confrontation systématique des données avec les règles proposées, d'où des avancées dans l'analyse. Des générateurs de paradigme ont également permis d'inclure des paradigmes exhaustifs dans les dictionnaires. Pour le dictionnaire khaling, une série de scripts Perl ont été écrits pour générer les paradigmes de conjugaison et convertir l'alphabet phonétique international en alphabet devanagari.

Pour aller plus avant dans la modélisation, nous nous orientons vers l'emploi de *transducteurs à états finis* : voir le traitement du yonaguni par Pellard & Yamada (sous presse). Pour traiter les effets

à longue distance de la morphophonologie en langue na (Michaud, 2017), nous prévoyons d'utiliser des langages spécialisés dans la programmation pour la linguistique (SLLP : *Specialized Languages for Linguistic Programming*).

4. Conclusion

Les dictionnaires de la collection Pangloss sont librement disponibles, et utilisables en l'état. Mais le rapide panorama présenté ici voulait surtout insister sur les possibilités qui s'ouvrent pour la suite du travail. On aimerait mentionner en conclusion le fait que la libre diffusion en ligne de *données connectées* dans des langues jusqu'ici pas ou peu dotées informatiquement (Berment, 2004) représente un enjeu soci(ét)al évident. Les dictionnaires de la collection Pangloss s'adressent à la fois aux linguistes et aux locuteurs des langues. Un soin particulier a été apporté au dictionnaire khaling, qui promeut une nouvelle orthographe distinguant toutes les oppositions phonologiques de cette langue et répondant aux exigences des locuteurs de la langue, contrairement à l'orthographe précédente, qui souffrait de nombreux défauts techniques. Au-delà de la conservation et de la mise à libre disposition d'un patrimoine culturel inestimable, des « tétralogies » connectées ouvrent de nombreuses possibilités telles que les environnements numériques personnalisés d'apprentissage (Mangeot, Bellynck, Eggers, Loiseau, & Goudin, 2016)

Remerciements

Nous sommes vivement reconnaissants envers les institutions et structures partenaires suivantes : ANR (projets Fondements Empiriques de la Linguistique, ANR-10-LABX-0083, et HimalCo, ANR-12-CORP-0006) ; CNRS-InSHS ; et Très Grande Infrastructure de Recherche Humanités Numériques (TGIR Huma-Num).

Références bibliographiques

Berment, V. (2004). *Méthodes pour informatiser les langues et les groupes de langues "peu dotées"* (thèse). Université Joseph Fourier - Grenoble 1.

Boas, F. (1902). *Tsimshian texts.* Washington : Government Printing Office.

Boas, F., & Swanton, J. R. (1911). Siouan (Dakota). In *Handbook of American Indian Languages I* (pp. 875–965). Washington : Government Printing Office, Bureau of American Ethnology, Bulletin 40.

Collins, S., Harrower, N., Haug, D. T. T., Immenhauser, B., Lauer, G., Orlandi, T., ... Wandl-Vogt, E. (2015). *Going Digital : Creating Change in the Humanities.* ALLEA.

Foley, W. A. (1999). Compte-rendu de Gerrit van Enk & Lourens de Vries, The Korowai of Irian Jaya. (Oxford studies in anthropological linguistics, 9). New York : Oxford University Press, 1997. Pp. xiv, 321. *Language in Society, 28*(3), 470–472.

François, A. (2003). *La sémantique du prédicat en mwotlap, Vanuatu.* Louvain : Peeters.

François, A. (2009). The languages of Vanikoro : Three lexicons and one grammar. In *Discovering history through language : papers in honour of Malcolm Ross* (pp. 103–126).

Francopoulo, G. (Ed.). (2013). *LMF : Lexical Markup Framework.* Wiley Online Library.

Jacobson, M., Michailovsky, B., & Lowe, J. B. (2001). Linguistic documents synchronizing sound and text. *Speech Communication, 33 [special issue : "Speech Annotation and Corpus Tools"]*, 79–96.

Jacques, G. (2015). Derivational verbal morphology in Khaling. *Bulletin of Chinese Linguistics, 8*(1), 78–85.

Jacques, G. (2016). Le sino-tibétain : polysynthétique ou isolant ? *Faits de Langues, 47*(1), 61–74.

Jacques, G., Lahaussois, A., Michailovsky, B., & Bahadur Rai, D. (2012). An overview of Khaling verbal morphology. *Language and Linguistics, 13*(6), 1095–1170.

Mangeot, M., Bellynck, V., Eggers, E., Loiseau, M., & Goudin, Y. (2016). Exploitation d'une base lexicale dans le cadre de la conception de l'ENPA Innovalangues. In *Enseignement des Langues et TAL* (Vol. 9, pp. 48–64). Paris : Association Francophone de la Communication Parlée.

Maxwell, M. (2012). Electronic grammars and reproducible research. In S. Nordhoff (Ed.), *Electronic Grammaticography* (pp. 207–235). Honolulu : University of Hawaii Press.

Michailovsky, B. (2002). *Limbu-English dictionary of the Mewa Khola dialect, with English-Limbu index.* Kathmandu : Mandala Book Point.

Michailovsky, B. (2011). Limbu. In D. Kouloughli & A. Peyraube (Eds.), *Encyclopédie des sciences du langage, Dictionnaire des langues* (pp. 1064–1074). Paris : Presses Universitaires de France.

Michailovsky, B., Mazaudon, M., Michaud, A., Guillaume, S., François, A., & Adamou, E. (2014). Documenting and researching endangered languages : the Pangloss Collection. *Language Documentation and Conservation, 8,* 119–135.

Michaud, A. (2017). *Tone in Yongning Na : lexical tones and morphotonology.* Berlin : Language Science Press.

Musgrave, S., & Thieberger, N. (2014, November). *Rethinking grammatical description : from Heath to hypertext.* Lecture, Research Unit for Indigenous Language, University of Melbourne. Disponible : https ://indiglang.arts.unimelb.edu.au/events grammatical-description-from-heath-to-hypertext/

Nordhoff, S. (2008). Electronic reference grammars for typology : challenges and solutions. *Language Documentation and Conservation, 2*(2), 296–324.

Pellard, T., & Yamada, M. (sous presse). Verb morphology and conjugation classes in Dunan (Yonaguni). In F. Kiefer, J. P. Blevins, & H. Bartos (Eds.), *Morphological paradigms and functions.* Leiden : Brill.

Romary, L. (2013). TEI and LMF crosswalks. *arXiv Preprint arXiv* :1301.2444.

Dealing with multiple orthographic standards within a single corpus: the case of Portuguese in the CoPEP corpus

Tanara Zingano Kuhn [1,2]*, José Pedro Ferreira [1], Maarten Janssen [1], Iztok Kosem [3,4], & Margarita Correia [1,2]

[1]CELGA-ILTEC, University of Coimbra, Portugal
[2]Faculty of Letters, University of Lisbon, Portugal
[3]Trojina, Institute for Applied Slovene Studies, Slovenia
[4]Faculty of Arts, University of Ljubljana, Slovenia

tanarazingano@yahoo.com, jpf@uc.pt, maartenpt@gmail.com, iztok.kosem@trojina.si, margarita@campus.ul.pt

1. Introduction

Corpora often have to deal with orthographic variation. In most cases, such variation is due to spelling errors. Apart from those intended for learners, most corpora choose to either ignore corrections and use only the original text, or ignore errors and use the corrected text instead. Both approaches have drawbacks. Multinational corpora of contemporary Portuguese such as CoPEP (see section 2) have to deal with the additional challenge of various spelling conventions at play, as prior to 2009 there were different conventions in Brazil and Portugal, and since then these were to a large extent merged through an international orthography treaty. The decision on which of the possible orthographic forms will be regarded as the most relevant depends on the purposes for which the corpus is being used. Besides, leaving three coexisting spellings unmarked in a single corpus can pose additional challenges to users. In this demonstration, we will show how CoPEP intends to deal with multiple orthographies using the TEITOK corpus framework. It should be stressed that, unlike widely adopted approaches with a focus on variant detection and normalisation (e.g., historical corpora), the perspective here does not involve standardisation, but rather a structured way for visualisation of diversity. It should be stressed that, unlike widely adopted approaches with a focus on variant detection and normalisation (e.g., historical corpora), the perspective here does not involve standardisation, but rather a structured way for visualisation of diversity, as will be briefly explained below.

2. CoPEP

CoPEP - *Corpus de Português Escrito em Periódicos* (Corpus of Written Portuguese in Academic Journals) (Kuhn & Ferreira, 2016) was compiled especially for the lexicographic project focussed on designing an online corpus-driven dictionary of Portuguese for university students (PhD research of the first author). The dictionary aims to describe the way the Portuguese language is used by expert writers from Brazil and Portugal in academic writing in different areas of knowledge.

CoPEP contains around 10,000 texts totalling over 40 million words extracted from academic journals published in the Brazilian and Portuguese national collections of SciELO (Scientific Electronic Library Online), distributed among three Schools of Knowledge, and further divided into six Great Areas. CoPEP is a synchronic corpus, the vast majority of its texts having been published between 2000 and 2016 (only 2% of texts are from the 1990s). The subcorpora for each language variety are of almost the same size and consist of a similar number of words per both Great Area and School, making the corpus evenly balanced.

Metadata on the texts have been carefully recorded in order to allow advanced corpus search options, e.g. year of publication, Great Area of Knowledge and ISSN number of a journal (the latter enabling interoperability with SciELO).

3. Orthographic variation in CoPEP

There are several different orthographic standards at play in CoPEP: the corpus consists of texts from both Brazil and Portugal written before and after 2009 when an international agreement on spelling (henceforth AO90) was signed. Brazilian texts published before 2009 follow the 1943 national spelling convention (henceforth AO43) (e.g. *idéia, freqüência, arquiinimigo,* now *ideia, frequência,*

∗. The first author is a Capes scholarhip holder, process number 0973/13-0.

arqui-inimigo); texts published in Portugal prior to 2009 adhere to the 1945 convention (henceforth AO45) (e.g. *acção, óptimo, neo-realismo,* today *ação, ótimo, neorrealismo*). Many forms were previously divergent and now have a common form, but for a number of words, the country-specific variation at a phonological level results in distinct orthographies even after the agreement reached in 2009 (e.g. *anônimo, recepção* in Brazil, but *anónimo, receção* in Portugal). Additionally, in a mostly phonemic orthography, there will always be variation between and within countries (e.g. *impacte* vs. *impacto*; *equipe* vs. *equipa*).

Such a multiplicity of orthographies poses some challenges when querying CoPEP and interferes with the collection of general statistics about the corpus. For instance, those strings that only differ due to codification differences must be searched each one at a time, e.g. language variety variants: *anónimo* (PT), *anônimo* (BR); spelling reform variants: *freqüente* (BR, pre-AO90), *frequente* (BR post-AO90 and PT, pre- and post-AO90), which in fact means the users must know that this variance exists in the first place. As to statistical results, each of those differently codified strings of characters whose meaning is the same is counted separately, making total frequency count skewed.

For data acquisition purposes, such as in the task for which the CoPEP corpus was originally developed, it would be useful to have variants (due to language variety, spelling reform or within-country variance) mapped so that a single query would yield results for all variants without losing track of the original spelling. Such results could be displayed in the original spelling, mapping into a single variant or following the AO90, the latter being ideal for the extraction of examples for pedagogical purposes.

4. Post-processing and visualising data

As Brazilian Portuguese and European Portuguese subcorpora can be created from CoPEP, the conversion of the texts from each variety to the current orthography involved running Lince, a tool for converting old spellings to new ones for both countries (Ferreira et al. 2012), on each subcorpus. For the task of linking other variant forms, we used the official word list for Portuguese, which contains explicit variant form linking (Ferreira, Correia & Almeida (Orgs.) 2017).

To model orthographic variants, CoPEP uses the TEITOK corpus framework, a corpus management system in which each document in a given corpus is represented by a separate XML document in the TEI/XML standard. In these XML files, tokens are modelled as in-line XML nodes. A searchable corpus is created automatically from all the XML files using Corpus WorkBench (Evert and Hardy, 2011, henceforth CWB). In TEITOK all orthographic variants are represented as a feature on those nodes. We can define as many orthographic forms as needed, for instance, *idéia/ideia* and *impacto/impacte* can be represented as follows in TEITO:

```
<tok corr= "idéia" ao90= "ideia">idea</tok>
<tok reg= "impacto">impacte</tok>
```

For most words, all these various spelling forms will be the same, which in TEITOK is solved by an inheritance mechanism: unless explicitly overwritten, each form will be identical to the form it depends on - the regularized orthography on the AO90 orthography, AO90 on the corrected spelling, and the corrected spelling on the written form. In the CWB corpus, all (implicit) forms are generated explicitly, meaning we can search for any of these forms, i.e. search for the regularised orthography, and see either the AO90 spelling or the original written form. The way this works in TEITOK is as follows: the user searches for the normalised orthography, e.g. [reg= "ideia"]. This will return all occurrences of *ideia* in the corpus, independently of how they were originally written in the source text. Rather than displaying the raw CWB output, TEITOK stores the document positions to look up the original XML, which contains all orthographic variants. In the interface, the users can then select the variant they are interested in (raw, post-AO90, variant-linked).

5. Concluding remarks

How orthography is dealt with in corpora depends on compilers' decisions, typically based on the purpose of the corpus. CoPEP was built as a basis for a dictionary for university students; thus

it should take normative standards into account. CoPEP on TEITOK holds multiple orthographies - AO90, AO43, and AO45; corrected misspelt words; and regularised variants - hence, it not only complies with its original purpose but also has its scope of use broadened, facilitating the interpretation of query results. One of the issues that still needs to be addressed is how to further tell apart different types of variants: *idéia* and *ideia* are clearly two forms of the same word, *impacte* and *impacto* less clearly so, and pairs of forms like estória and história may no longer be used as variants of each other, although they started out as such.

References

Ferreira, J. P., Correia, M. & Almeida, G. B. de (orgs.) (2017) *Vocabulário Ortográfico Comum da Língua Portuguesa.* Praia: IILP. Available at: http://voc.cplp.org

Evert, S. & Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. *Proceedings of the Corpus Linguistics 2011 conference, University of Birmingham, UK.* [PDF] Available at http://www.stefan-evert.de/PUB/EvertHardie2011.pdf (Accessed on 30 January 2017).

Ferreira, J.P., Lourinho, A. & Correia, M.. (2012). Lince, an End User Tool for the Implementation of the Spelling Reform of Portuguese. In Helena Caseli et al. (eds.). *Computational Processing of the Portuguese Language,* 46-55. Berlin, Heidelberg: Springer.

Janssen, M. (2016). TEITOK: Text-Faithful Annotated Corpora. *Proceedings of LREC 2016, Portorož, Slovenia,* 4037-4043.

Kuhn, T.Z., & Ferreira, J.P. (2016). Building a corpus of written academic texts in Portuguese. *Teaching and Language Corpora Conference (TaLC12), Giessen, Germany. Book of Abstracts,* 103. Giessen.

Scielo Brazil. Available at: www.scielo.br (Accessed on 15 February 2016).

Scielo Portugal. Available at: www.scielo.mec.pt (Accessed on 1 February 2016).

Session 3.B.
Énonciation et registre

Usages des adverbes de domaine : Analyse automatisée du profil combinatoire

Dennis Wandel
Université de Neuchâtel – Suisse
dennis.wandel@unine.ch

1. Adverbes de domaine : Organisateurs de l'énonciation ?

L'analyse quantitative des adverbes épistémiques (*certes, peut-être, sans doute, certainement, sûrement...*) menée dans Rossari, Hütsch, Ricci, Salsmann, et Wandel (2016) montre que certains de ces adverbes sont régulièrement utilisés dans le premier membre d'une structure concessive construite avec le connecteur *mais*. Selon Rossari (à paraître-a, à paraître-b), la lecture concessive est issue d'un contraste entre deux énonciations que *mais* déclenche. En cooccurrence en tant que collocatif gauche de *mais*, ces formes ont tendance à renforcer ce contraste énonciatif. Nous faisons l'hypothèse que des formes comme *en théorie, en apparence, en principe, en réalité, en vérité, en fait* et *en pratique* peuvent également souligner un tel contraste. En plus de leur fonction d'adverbial de domaine (Féron, 2007 ; Guimier, 2007 ; Raemdonck, 1999), elles peuvent endosser une fonction au niveau énonciatif en donnant des indications sur le type de prise en charge d'un énoncé. Dans cet usage, leur fonction cadrative (liée à leur emploi d'adverbe de domaine) se reporte sur l'énonciation, permettant ainsi de contraster subtilement la prise en charge (Dendale & Coltier, 2011) de P par rapport à celle de Q.

> (P) « On met tout sur la personnalité de Fillon, ***mais*** (Q) ***en réalité***, c'est la fonction de premier ministre qui est dépréciée au fil du temps. Fillon l'avait théorisé et il est le premier à le subir », analyse- t-on à l'Elysée. *(BTLC.Primestat – Le Monde 2008)*

Pour tester notre hypothèse, nous utilisons les outils statistiques. Au moyen du concept du *profil combinatoire* (Blumenthal, 2008) qui se sert des calculs de spécificité pour mesurer la spécificité des cooccurrences de mots, nous vérifierons si ces adverbes interviennent en fait typiquement dans une concession où ils paraissent susceptibles d'une fonction énonciative.

2. Profils combinatoires des adverbes de domaine : une analyse sur corpus des usages dans les contextes typiques

L'objectif est d'expliquer les usages concessifs des organisateurs de l'énonciation au moyen d'une analyse menée sur des corpus du genre textuel pour déterminer dans quels contextes ces adverbes sont typiquement utilisés (avec fonction cadrative et énonciative). Blumenthal (2008) ; Blumenthal, Diwersy, et Mielebacher (2005) ont décrit la méthodologie derrière notre analyse outillée de l'usage : L'extraction automatisée des profils combinatoires d'un mot pivot – en d'autres termes l'ensemble des combinaisons statistiquement spécifiques avec leurs mots cooccurrents – de chaque organisateur de l'énonciation au moyen des mesures d'association statistiques (notamment *log-likelihood*) permet par la suite de comparer les contextes dans lesquels interviennent ces adverbes. Cette démarche méthodologique a l'avantage d'être inductive. Les contextes spécifiques des adverbes peuvent être extraites sans les connaître tous connaître auparavant. Une contrainte de cette approche méthodologique est que l'analyse automatisée sur corpus ne peut pas différencier entre fonction cadrative et fonction énonciative des adverbes en raison d'absence d'une annotation sémantique qui distinguerait toutes les nuances sémantiques d'un mot. Une évaluation qualitative des résultats quantitatifs s'avère indispensable. La recherche quantitative des profils combinatoires des adverbes permet d'analyser deux usages des organisateurs de l'énonciation qui sont d'intérêt particulier pour notre analyse quantitative.

Premièrement, il faut déterminer si l'on trouve les adverbes de domaine typiquement dans une structure concessive avec *mais* à l'instar des adverbes épistémiques (*certes*, etc.).

Deuxièmement, afin d'analyser les facteurs qui possiblement facilitent un fonctionnement énonciatif des adverbes, nous cherchons à savoir si les adverbes de domaine interviennent dans les mêmes contextes et si deux adverbes peuvent aussi être fréquents dans la même structure.

Dans l'analyse des résultats quantitatifs, il est question de savoir quelles fonctions (cadratives ou/et énonciatives) ces adverbes ont dans ces usages concessifs. Nous supposons trouver des usages cadratifs ainsi qu'énonciatifs. En ce qui concerne les énonciatifs, nous avons l'hypothèse qu'il existe deux types d'usages énonciatifs : un premier groupe d'adverbes qui représente ce qui est mis en retrait dans l'argumentation (*en théorie, en apparence, en principe*) et un second groupe d'adverbes qui marque ce qui est mis en avant (*en réalité, en vérité, en pratique, en fait*).

2.1. Extraction automatisée du profil combinatoire

Les mesures d'association statistiques permettent de trouver les combinaisons spécifiques (cooccurrences) des mots de façon automatisée. Ces calculs sont probabilistes et se basent uniquement sur les fréquences des mots dans un corpus et ainsi ne prennent en compte que les aspects sémantiques. Pour nos requêtes automatisées, nous utilisons la plateforme *BTLC.Primestat*. Les adverbes de domaine sont définis comme mots pivots de six recherches indépendantes. En cherchant les mot-collocatifs spécifiques au moyen du calcul de *log-likelihood*, on obtient sept profils combinatoires pour ces sept adverbes. Tableau 1 récapitule seulement les collocatifs spécifiques que nous considérons comme indicateurs d'un usage concessif (notamment le connecteur *mais*) ou d'un usage énonciatif en cooccurrence avec un autre adverbe du même type.

Le corpus du journal *Le Monde* de 2008 est le seul corpus de *BTLC.Primestat* qui permet de chercher une préposition suivie par un nom (*en + nom*) et comme pivot et comme collocatif. Les autres corpus disponibles ne permettent pas de chercher *en + nom* en tant que collocatif et compliquent ainsi une recherche représentative. Une solution limitée mais pratiquable est une recherche basée sur la séquence exacte (*en + théorie + .* + en + réalité*) afin de vérifier si des combinaisons de deux adverbes de domaine sont spécifiques. En revanche, ceci ne permet pas d'établir automatiquement l'ensemble du profil combinatoire de l'adverbe.

2.2. Résultat 1 : mais comme cooccurrent spécifique des adverbes de domaine

Le premier résultat de l'analyse quantitative du profil combinatoire est que le marqueur concessif prototypique [1] *mais* est parmi les collocatifs spécifiques de tous ces adverbes analysés. Vu le cas des adverbes épistémiques qui prennent une fonction énonciative en tant que cooccurrent spécifique de *mais*, ce résultat de notre recherche est un indice que les adverbes de domaine ont aussi un potentiel énonciatif spécifique dans une concession avec *mais*. À partir des exemples de cooccurrences sorties du concordancier de *BTLC.Primestat*, il est possible d'attester et des emplois cadratifs et des emplois énonciatifs qui ressemblent ceux des adverbes épistémiques.

2.3. Résultat 2 : Combinaisons spécifiques d'adverbes de domaine

Presque tous les adverbes sont spécifiques avec au moins un autre adverbe (sauf *en vérité*) ou une variante calquée sur le même nom (*dans la théorie,* etc.). *(En) théorie* et *(en) pratique* sont considérés comme collocatif spécifique de l'autre selon le *log-likelihood*. De même pour *(en) apparence* et *(en) réalité*. *Réalité* est aussi spécifique de en principe. *Vérité* est spécifique de *en apparence*. *Réalité* est spécifique de *en principe*. On constate une certaine régularité dans la combinaison des adverbes quant au cadre ou domaine qu'ils représentent. Deux types d'adverbes se manifestent : un marqueur moins fort et plutôt abstrait (*en théorie, en apparence, en principe*) se combine souvent avec un marqueur plus fort et plutôt factuel (*en vérité, en pratique, en réalité*).

Nous pouvons attester des usages cadratifs des adverbes et, moins fréquemment, des usages énonciatifs qui ne désignent aucune restriction à un domaine sémantique, mais une qualification au niveau de l'énonciation qui représente le type de prise en charge du locuteur. Par conséquent, dans ces deux emplois, cadratif et énonciatif, on ne trouve spécifiquement ensemble ni *en théorie*, *en apparence* et *en principe* ni *en vérité*, *en pratique*, *en fait* et *en réalité*, mais seulement des combinaisons d'adverbes de

1. Nous avons comparé les fréquences des marqueurs concessifs dans quatre corpus. Mais est largement plus fréquent que les autres marqueurs concessifs (*pourtant, toutefois, cependant, néanmoins*) (voir tableau 2). Nous le considérons comme marqueur prototypique d'une concession.

type différent. Nous supposons donc deux groupes d'adverbes au fonctionnement énonciatif qui chacun ne peuvent représenter qu'un type de prise en charge dans une concession. Les (co-)occurrrences représentées dans les résultats de l'analyse automatisée sur corpus semblent confirmer cette hypothèse.

3. Limites et perspectives de l'analyse automatisée sur corpus des adverbes à fonctionnement cadratif et énonciatif.

Maquant une annotation sémantique, surtout pour les nuances sémantiques d'un mot, telle que la fonction énonciative, qui ne sont pas dénotationnelles et parfois floues, la requête automatisée sur corpus offre néanmoins une nouvelle perspective quantitative sur les usages des adverbes de domaine. En se basant sur les corpus, l'analyse des profils combinatoires permet d'expliquer la synergie de l'usage et des fonctions énonciatives d'un mot au moyen des facteurs statistiques et de manière systématique. Mais cette analyse nécessite impérativement une évaluation qualitative et un tri manuel des résultats pour l'interprétation des fonctions énonciatives. Concrètement, il est nécessaire de décider pour chaque occurrence d'un adverbe de domaine s'il s'agit d'un emploi cadratif ou énonciatif. À la base de ce tri manuel, il sera possible de se prononcer sur la proportionnalité des emplois cadratifs et énonciatifs. Reste que l'analyse quantitative n'offre que des indications pour analyser les facteurs qui permettent un fonctionnement énonciatif des adverbes de domaine (cooccurrence avec *mais*, type de prise en charge, mais aussi noyau sémantique, position syntaxique, contexte, genre textuel, etc.).

4. Comparaison des profils combinatoires dans corpus de différents genres

Les résultats présentés sont issus de l'analyse d'un corpus de presse (*BTLC.Primestat – Le Monde 2008*). Dans un deuxième tour, il faudra refaire l'analyse automatisée des profils combinatoires pour trois autres corpus comparables du genre journalistique (*Le Monde 2007, Le Figaro 2007, Le Figaro 2008*) afin de déterminer la représentativité des résultats tiré du corpus choisi et pour comparer les profils combinatoires extraits. Dans un troisième tour, des corpus de trois autres genres textuels seront introduits dans une analyse de corpus contrastive. Nous disposons de deux corpus encyclopédiques, l'Encyclopédie de Diderot et d'Alembert (*ARTFL*) et l'encyclopédie *Universalis* (*Encyclopaedia Universalis*), d'un corpus encyclopédique en ligne (*Wikipédia français*), d'un corpus Web (*French Web 2012 (frTenTen12)*) et d'un corpus du genre discursif écrit (*Europarl*). Nous nous attendons à une plus grande fréquence des adverbes de domaine en tant qu'organisateurs de l'énonciation dans les textes argumentatifs du genre de presse et dans les discours politiques que dans les encyclopédies et sur le Web auxquels nous associons un style plus neutre. Il reste à voir si les organisateurs de l'énonciation sont également spécifiques du corpus *French Web 2012*, ce dernier étant le plus hétérogène de ces corpus.

5. Tableaux et graphiques

5.1. Tableau 1 : Liste des collocatifs spécifiques des adverbes de domaine (Le Monde 08 – récapitulatif des sept profils combinatoires établis)

Mot pivot	Mot cooccurrent	Contexte gauche (L) ou droit (R)	*log-likelihood ll*	Rang *log-likelihood*	Fréquence de cooccurrence
En théorie	***En pratique***	R	45.21	1.	5
En théorie	***Mais***	R	17.76	4.	20
En apparence	***En réalité***	R	22.52	16.	4
En apparence	***En fait***	R	19.66	23.	4
En apparence	***Mais***	R	8.62	385.	15
En apparence	***En vérité***	R	7.23	542.	1
En réalité	***Mais***	L	49.64	2.	88
En réalité	***En apparence***	L	22.49	20.	4
En réalité	***En principe***	L	11.18	341.	3
En fait	***En apparence***	L	26.72	17.	5
En fait	***Mais***	L	21.26	39.	91
En fait	***En principe***	L	9.15	233.	3
En vérité	***Mais***	L	18.07	34.	15
En pratique	***En théorie***	R	44.91	3.	5
En pratique	***Mais***	L	14.65	97.	21
En pratique	***En principe***	L	4.57	979.	1
En principe	***Mais***	R	22.74	10.	33
En principe	***En réalité***	R	11.23	219.	3
En principe	***En fait***	R	9.18	406.	3
En principe	***En pratique***	R	4.58	1256.	1

TABLE 1 – Liste des collocatifs spécifiques des adverbes de domaine (Le Monde 08 – récapitulatif des sept profils combinatoires établis)

5.2. Tableau 2 : Fréquences relatives des marqueurs concessifs dans quatre corpus

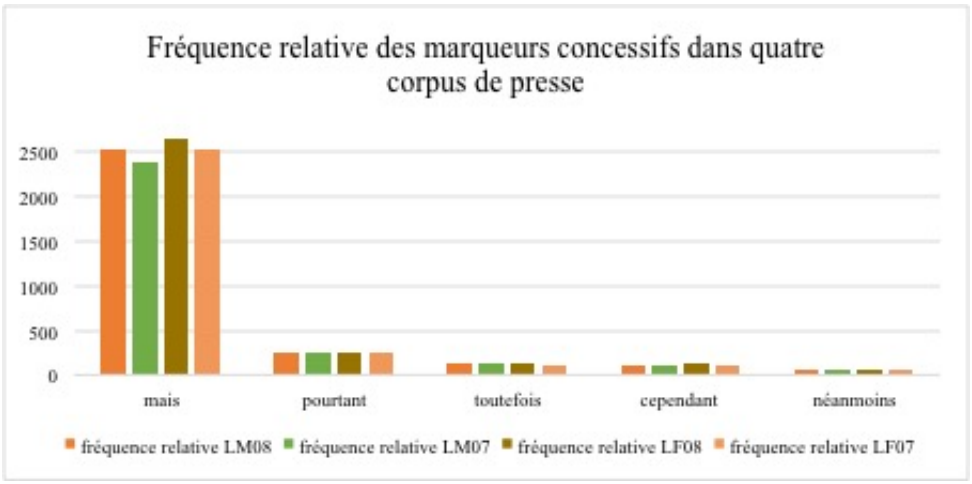


TABLE 2 – Fréquences relatives des marqueurs concessifs dans quatre corpus

1. LM 07/08 = *Le Monde 2007/2008* ; LF 07/08 = *Le Figaro 2007/2008*

Références bibliographiques

Blumenthal, P. (2008). Combinatoire des prépositions : approche quantitative. *Langue française, 157,* 37-51. doi :10.3917/lf.157.0037

Blumenthal, P., Diwersy, S., & Mielebacher, J. (2005). Kombinatorische wortprofile und profilkontraste. Berechnungsverfahren und Anwendungen. *Zeitschrift für romanische Philologie, 121,* 49-83.

Dendale, P., & Coltier, D. (2011). *La prise en charge énonciative : études théoriques et empiriques.* Bruxelles : De Boeck Supérieur.

Féron, C. (2007). Les adverbiaux *en vérité* et *à la vérité* en moyen français et en Français préclassique. In B. Combettes & C. Marchello-Nizia (Eds.), *Études sur le changement linguistique en français* (pp. 145-156). Nancy : Presses Universitaires.

Guimier, C. (2007). Adverbe de domaine et structuration du discours. In A. Celle, S. Gresset, & R. Huart (Eds.), *Les connecteurs, jalons du discours* (Vol. Sciences pour la communication pp. 43-70). Bern : Peter Lang.

Raemdonck, V. (1999). L'adverbe de domaine-point de vue est-il un adverbe de phrase ? *Orbis Linguarum, 11,* 101-112.

Rossari, C. (à paraître-a). L'approbation dans un dialogue devient-elle une concession dans un monologue ? Etude de certes, en effet, effectivement, d'accord, OK. In L. Sarda, D. Vigier, & B. Combettes (Eds.), *Mélanges pour Michel Charolles. Connexion et indexation. Ces liens qui tissent le texte.* Paris : Armand Colin

Rossari, C. (à paraître-b). La concession sans opposition à la lumière de la théorie argumentative de la polyphonie *Verbum, 38 : Théorie des blocs sémantiques et Théorie argumentative de la polyphonie.*

Rossari, C., Hütsch, A., Ricci, C., Salsmann, M., & Wandel, D. (2016). Le pouvoir attracteur de mais sur le paradigme des adverbes épistémiques : du quantitatif au qualitatif. Paper presented at the *Actes de Colloques JADT 13èmes Journées internationales d'Analyse statistique des Données Textuelles, Nice.*

The Role of Grammaticalization in Interweaving the Personal and Impersonal in Stance Expression in Formal Spoken English

Amira Agameya
The American University in Cairo

Abstract

The expression of speaker/writer stance is a fundamental feature of intersubjective positioning both in speaking and writing, where the speaker/writer conveys a given point of view (Hyland, 2005 ; White, 2003). Traditionally, formal written English required the use of an impersonal, detached mode of communication by employing a range of lexico-grammatical devices to ensure objectivity. Some of these devices include constructions such as existential *there*, the *it ... that* structure and the passive voice. These devices allow the writer to express an opinion or attitude implicitly without resorting to self-mention (Berman et al, 2002 ; Berman, 2005 ; Biber & Finegan, 1989). In informal spoken language, by contrast, reference to the speaker is expected and hedging devices typically include expressions using the first-person pronoun, e.g. *I think*. Formal spoken language exhibits features from both. Focusing on one particular grammatical structure, *it ... that*, associated with impersonal writer stance in academic writing, the present corpus-based study examines its use in current American formal spoken English in an attempt to address three issues having to do with (a) the role of interpersonal relations in speech, (b) the role of grammaticalization in assigning new functions to a grammatical construction, and (c) the nature of the interaction between interpersonal expressions and the grammaticalized construction. The corpora used for the study are COCA (Corpus of Contemporary American English) and CSPAE (Corpus of Spoken Professional American English) as well as COHA (Corpus of Historical American English) for comparison purposes. The study hypothesizes that, for pragmatic reasons, the nature of intersubjective positioning in spoken interaction requires the use of discourse markers and hedging (White, 2003), while the nature of the formal context calls for using a formal construction, e.g. *it ... that*, which overtime has been grammaticalized so that it became specialized in expressing formal stance. As a consequence of grammaticalization, it has almost lost its impersonal nature, at least in formal speech.

Based on the analysis of the CSPAE and the spoken section of COCA and, particularly structures such as *I think it 's quite clear that it's important that ; I mean, it seems to me that ; I mean, honestly, it 's very clear that ; I mean, it's not surprising that ; I was gonna say, you know, yes, it's also true that ; And I think it 's pretty clear that ; I believe it is very important that ; I think it's a shame that ; I mean/think, it's very interesting/no surprise that*, I argue that the nature of the formal discourse context seems to impose the impersonal structure on the speaker and, at the same time, the pragmatics of spoken interaction requires the speaker to use hedges, among other things, for engagement, or rapport management purposes (Spencer-Oatey, 2005). This, I will contend, has led the use of this impersonal structure to be conventionalized in the spoken variety, perhaps through the process of pragmatic strengthening (Traugott, 1989). Rule generalization, in turn, resulted in expanding the contexts in which this grammatical pattern is used from writing to speaking (Hopper and Traugott, 2003). The result of expansion has led the structure to be associated with formal expression of stance in general, without the corollary feature of necessarily being impersonal. In other words, the structure has been grammaticalized and, following (Diewald, 2011), pargmaticalized, i.e. it has come to perform a new pragmatic function associated with formality, where pragmaticalization involves speaker's attitude towards the hearer (Aijmer, 1997).

References

Aijmer, K. (1997) *I think*—an English modal particle. In Toril Swan & Olaf Jansen Westvik (eds.), *Modality in Germanic languages. Historical and comparative perspectives* (Trends in Linguistics : Studies and Monographs 99), 1– 47. Berlin & New York : Mouton de Gruyter.

Berman, R. (2005). Introduction : Developing discourse stance in different text types and languages. *Journal of Pragmatics* 37, 105-124.

Berman, R., Ragnarsdottir, H. and Stromqvist, S. (2002). Discourse stance. *Written Language and Literacy* 5, 255-290.

Biber, D. and Finegan, D. (1989). Styles of stance in English : Lexical and grammatical marking of evidentiality and affect. *Text* 9 :1, 93-124.

Diewald, G. (2011). Pragmaticalization (defined) as grammaticalization. *Linguistics* 49 :2, 365-390.

Hopper, P. J., and Traugott, E. 2003. Grammaticalization. 2nd ed. Cambridge : Cambridge University Press.

Hyland, K. (2005). Stance and engagement : A model of interaction in academic discourse. *Discourse Studies* 7 :2, 173-192l

Spencer-Oatey, Helen, 2005. (Im)Politeness, face and perceptions of rapport : unpackaging their bases and interrelationships. *Journal of Politeness Research* 1 :1, 95–119.

Traugott, E. (1989). On the rise of epistemic meanings in English : An example of subjectification in semantic change. *Language* 65(1). 31–55.

White, P. P. R. (2003). Beyond modality and hedging : A dialogic view of the language of intersubjective stance. *Text* 23 :2, 259-284.

Session 4.A.
Phraséologie

Routines conversationnelles dans le roman policier : interrogatoire

Teresa Muryn et Małgorzata Niziołek
Université Pédagogique de Cracovie
teresa.muryn@gmail.com, mniziolek1@gmail.com

1. Introduction

L'objectif de cette étude est d'analyser les routines/formules associées à quelques situations conversationnelles spécifiques dans les romans policiers. Il s'agit pour nous de repérer les structures récurrentes qui apparaissent aux différentes étapes d'un interrogatoire policier. Le corpus sur lequel se fonde cette étude se compose de romans policiers en français venant d'auteurs reconnus, aussi bien français qu'étrangers et compte actuellement 35850957 tokens. Dans plusieurs situations le comportement langagier des interlocuteurs est soumis à certaines restrictions, souvent considérables. Les interlocuteurs, les participants à un dialogue, jouent des rôles et respectent des schémas bien définis préalablement. Les situations dans lesquelles se trouvent les locuteurs les contraignent à employer telle ou telle structure. Nous commencerons par rappeler la définition des routines conversationnelles pour passer ensuite à la présentation des structures extraites du corpus et proposer leur classement.

2. Quelques remarques sur la notion de routine conversationnelle

La notion de routine discursive est assez floue. Plusieurs études ont été consacrées à ce type d'énoncés, appelés maximes conversationnelles (Grice), routines conversationnelles (Traverso, Klein et Lamiroy), routines/formules discursives, énoncés formulaires (Sfar, Schapira). Selon Traverso (2006 : 41) les routines présentent les caractéristiques suivantes : ce sont des formules préfabriquées, partiellement ou totalement vides de contenus sémantiques, adaptées à une situation particulière et partagées par l'ensemble des membres d'une société. Cependant, les unités que nous avons récupérées ne sont pas sémantiquement opaques. Elles ne présentent pas non plus de déviations syntaxiques qui pourraient démontrer leur caractère figé. En plus le sens de ces séquences est compositionnel. Ces formules sont associées à une situation de communication spécifique, qui impose l'emploi d'expressions précises au lieu d'autres, similaires, qui pourraient sémantiquement convenir, mais non pragmatiquement. Sfar (2007) insiste sur le caractère figé (voir Mejri, 1997, Gross 1996) de ces unités, mais propose un classement qui prend en compte aussi bien des énoncés qui répondent aux critères du figement que ceux qui acceptent des transformations (substitution synonymique et l'insertion d'éléments facultatifs) (Sfar, 2007 : 318), Niziołek s'intéresse, à son tour, aux fonctions des routines conversationnelles dans le texte littéraire (2013) .

3. Routines conversationnelles dans la scène d'interrogatoire

Il faut souligner que les structures que nous voulons analyser connaissent souvent des variantes. Ces routines discursives repérées acceptent des transformations et leur structure peut être modifiée. Pendant un interrogatoire les policiers emploient souvent un répertoire restreint de formules "toutes faites", de plus, la structure séquentielle est plus ou moins réglée (et prévisible). Des suspects recourent aussi aux séquences routinières. Il existe pour les interrogatoires une forme de déroulement relativement figée, elle comporte des activités quasiment obligatoires. Ces activités se réalisent dans l'emploi des structures préfabriquées qui ont un contenu pragmatique bien déterminé. Traverso (2006 : 26) remarque que «les routines favorisent la mise en ordre de l'interaction », ce qui est particulièrement souhaitable lorsque l'on mène une enquête. L'emploi des routines conversationnelles indique le/les rôle(s) des locuteurs : le policier, le coupable, le témoin. L'emploi de certaines séquences permet de les identifier, ou au moins d'identifier les fonctions qu'ils assument (C'est parce que chacun joue un rôle qui lui revient de façon répétitive). Certains milieux sont facilement identifiables par le langage employé. Même les réponses des coupables s'inscrivent dans cette logique situationnelle. C'est ainsi également que l'individu construit son « identité situationnelle ». Les routines deviennent un signe d'appartenance à un groupe, une manière de prendre les rôles interactionnels attachés à un type de situation. Même

si la langue a un caractère créatif, le fait de participer aux mêmes situations nous fait recourir à des séquences préfabriquées.

Enfin, il ne faut pas oublier que l'interrogatoire policier n'est en rien une conversation. C'est un condensé de tentatives de persuasion, de tactiques argumentatives, qui s'adaptent au contexte et à la personnalité des interrogés ; ce qui explique les variations dans les formules utilisées. Même si ce sont des romans policiers, construits sur des modèles bien décrits, ce sont surtout des textes littéraires dans lesquels une liberté « productive » est envisageable et même désirable. Cela peut expliquer des « variations » dans la structure des formules repérées. Mais la présence de « variantes » des routines discursives peut trouver encore une explication. Il ne faut pas considérer ces variantes comme des écarts par rapport à une liste fermée de formules figées. Au sein de la même formule, dont l'emploi est imposé par la profession qu'on exerce (ici, l'enquêteur), on peut retrouver quelques variantes « fonctionnelles anaphoriques » qui ne déforment pas la routine-source. Le contexte reste le même, les rôles des interlocuteurs également. C'est un procédé qui permet de personnaliser ces formules. Il reste à savoir si on a le droit de les ranger parmi les routines conversationnelles.

Au niveau sémantique, il existe trois concepts propres au genre policier : *personnage-crime-enquête*. Le motif sémantique est un schéma de concepts obligatoires se réalisant dans une situation précise. Le motif de l'interrogatoire fait partie de l'enquête et il se compose des éléments suivants :

a. personne chargée de l'enquête + lieu : un commissariat + suspect/témoin + crime

b. personne chargée de l'enquête + lieu : autre que commissariat + personne impliquée + crime

Ci-dessous nous présentons les premiers résultats extraits du corpus. L'accès aux routines a été possible grâce aux mots clés, entre autres, verbes de parole (surtout verbes introducteurs du discours) et structures lexico-syntaxiques introductrices à la scène de l'interrogatoire.

a) Introduction- Présentation d'un officier de police

- *Commissaire X, de la Police judiciaire (de + ville).*

b) Entrée en contact

- *Vous me permettez de vous poser quelques questions ?*
- *J'appartiens à la Police Judiciaire et je voudrais vous poser quelques questions...*
- *J'ai simplement quelques questions à vous poser...*

c) Identification de la personne interrogée

- *Comment vous appelez-vous ?* (identité)
- *Vous avez une profession ?* (profession)
- *Que faisiez-vous auparavant ?/ Qu'est-ce que vous faisiez auparavant ?* (profession)
- *Inscrite au registre de la police des mœurs ?*(expérience)

d) Relation entre la personne interrogée et la victime

- *Voulez-vous, monsieur X, regarder cette photographie et me dire si c'est bien celle de l'homme que vous avez reçu ici une nuit de la semaine dernière ?... Il tendit un portrait du mort.*
- *Avez-vous eu des discussions avec Nhum (la victime) ?*
- *Quand avez-vous appris la mort de votre Nhum ?*

e) Questions sur l'alibi :
I. Heure

- *Quelle heure était-il alors ?*
- *Savez-vous quelle heure il était ?*

- *Pourriez-vous me dire à quelle heure il vous a quitté ?*
- *Voulez-vous me dire à quelle heure vous avez vu Nhum pour la dernière fois ?*
- *À quelle heure êtes-vous rentré dans votre chambre ?*

II. Lieu

- *Où étiez-vous (+ la date/heure/moment de la journée) ?*

III. Témoin

- *Nhum (Nom Propre) vous accompagnait ?*
- *Vous l'avez donc vu ?*

e) Questions concernant le suspect
Identification de la personne soupçonnée :

- *Est-ce que vous connaissez cette personne ?*
- *Vous la/le reconnaissez ? / Vous reconnaîtriez l'homme ?*
- *Vous connaissez Nhum ?*
- *Tout ce que je voudrais savoir, c'est ce que vous pensez de Nhum*
- *Il était habillé en N ?*

Au cours d'interrogatoire :

- *Vous n'avez rien de particulier à m'apprendre ?*
- *Me permettez-vous quelques questions indiscrètes ?...*
- *Alors, j'ai besoin que vous me confirmiez la vérité...*

Fin d'interrogatoire :

- *Maintenez-vous cette affirmation ?*
- *Vous maintenez toutes vos déclarations ?*
- *Vous n'avez rien à ajouter ?*
- *Dans ce cas, il ne me reste qu'à vous remercier et à m'excuser encore, monsieur + Nom Propre...*

4. Perspectives

La reconnaissance des routines conversationnelles permet de situer chronologiquement chaque scène dans le roman policier. Chaque routine ne peut apparaître que dans une situation réelle spécifique. Elle perd sa fonction en dehors de cette situation. C'est ainsi que l'emploi d'une routine joue le rôle discriminant pour une situation décrite parce qu'elle implique ses circonstances. L'extraction des routines et leur classification mènent à la découverte de l'organisation interne du schéma linéaire du roman policier (même si ce n'est pas son seul critère)

Références bibliographiques

François, M. (2009). « Le stéréotype dans le roman policier ». *Cahiers de Narratologie,* [En ligne], 17 | 2009, mis en ligne le 22 décembre 2009, consulté le 30 mars 2015. URL : http://narratologie.revues.org/1095.

Gross, G. (1996). *Les expressions figées en français. Noms composés et autres locutions.* Paris : Ophrys.

Klein J. R., Lamiroy B., , 2011). Routines conversationnelles et figement. (éds) Anscombre, J-C.et Salah, M., *Le figement linguistique : la parole entravée,*, 195-213.

Mejri, S. (2011). *Le figement linguistique : la parole entravée.* Honoré Champion.

Lüger H.-H. (1993). Routine conversationnelle et comportement langagier. *Langage et société,* n/upo63, 5-38.

Muryn T., Niziołek M., Hajok A., Prażuch W., Gabrysiak K. (2016). Scène de crime dans le roman policier : essai d'analyse lexico-syntaxique. *Actes du CMLF2016 : http://dx.doi.org/10.1051/shsconf/20162706007.*

Muryn, T. Niziołek, M. (2016). Pour une analyse phraséologique du roman policier. (éds Mogorron Huetra, Pedro, Cuadrado Rey, Analia, Martinez Blasco, Iván, Navarro Brotons, Lucia. *Fraseologia, variacion y traducción,* Peter Lang, 139-151.

Niziołek, M. (2013). Étude contrastive des routines discursives (conversationnelles) dans le roman policier : l'exemple des romans de Georges Simenon (la série « Maigret »). (éds.) Muryn, T., Mejri, S. & all. *La phraséologie entre langues et cultures.* PeterLang. 161-170.

Sfar, I. (2007). Les énoncés formulaires : contenu pragmatique et problèmes de traduction, *A la croisée des mots. Hommages Taïeb Baccouche,* Université de Sousse ; Université Paris 13. Sousse ; Villetaneuse, 2007. 313-328

Traverso, V. (2006). *Des échanges ordinaires à Damas : aspects de l'interaction en arabe. Approche comparative et interculturelle,* Lyon, PUL.

Séquences récurrentes typiques du roman policier et de science-fiction : une étude préliminaire sur corpus

Judith Chambre et Olivier Kraif
LIDILEM, Univ. Grenoble Alpes
Judith.Chambre@gmail.com, Olivier.Kraif@univ-grenoble-alpes.fr

1. Introduction

La littérature de genre, ou paralittérature, est considérée comme étant très codifiée. Selon Boyer, « chaque genre comprend un certain nombre de sous-ensembles, des séries fondées sur la réutilisation de composantes identiques : réapparition d'un même protagoniste ou, du moins, recours à un typologie à peu près fixe de personnages ; fonctions déterminées attribuées à chacun d'entre eux ; constantes stylistiques ; utilisation d'un décor propre ; techniques spécifiques d'agencement des épisodes ; etc. » (Boyer, 1992, p.97) Ces éléments permettent d'inscrire le récit dans un genre. Ainsi la spécificité de chaque sous-genre est déterminée par une dominante qui attribue une place et un rôle à chaque personnage, action (Ibid.).

La récurrence de ces éléments dans les livres d'un même genre contribue à forger un « horizon d'attente » chez le lecteur. De la sorte, un lecteur qui lit un polar s'attend à retrouver certains types de personnages, de situations. Une sorte de contrat de lecture est passé entre l'auteur et le lecteur (Boyer, 1992) : il permet d'orienter le lecteur et lui assure que le récit qu'il lit sera conforme à ce qu'il attend. Le titre du livre ainsi que la couverture sont également des composants de ce cadre de lecture. Ainsi, que ce soit au niveau de la structure du texte ou de son contenu, des éléments se retrouvent dans tous les textes appartenant à un genre. Si l'on compare différents sous-genres, tels que la science-fiction, le roman policier, la fantasy, chacun véhicule donc des spécificités et des normes différentes.

Selon Siepmann (2015 ; 2016), la littérarité d'un texte se fonde sur la sur-représentation significative de mots et de motifs. Se basant sur l'hypothèse selon laquelle il est possible d'identifier un genre grâce à différents indices, des études ont été menées pour étudier les constructions spécifiques à un genre, comme la littérature sentimentale (Legallois, Charnois, & Poibeau, 2016) ou le roman policier et la science-fiction (Kraif, Novakova, & Sorba, 2016).

Notre étude, conduite dans le cadre du projet PhraseoRom (financement ANR-DFG) qui vise à étudier la phraséologie du genre romanesque et de ses sous-genres, s'inscrit dans la même perspective. Plus précisément, notre objectif est d'identifier et de comparer les traits spécifiques à deux sous-genres, le roman policier et la science-fiction, en anglais et en français. Différents éléments peuvent être pris en compte pour comparer les genres : le vocabulaire, les séquences récurrentes, les constructions syntaxiques, ou des critères textométriques tels que les longueurs des phrases ou le rapport type/occurrence. Ainsi Biber et Conrad (2009) et Biber (2012) ont montré que des critères linguistiques variés peuvent être pertinents dans la comparaison des registres et des discours : la proportion de verbes, pronoms et noms, l'utilisation de certains pronoms, les temps des verbes, les tournures à la voix passives, les subordonnées complétives et relatives, les syntagmes prépositionnels complément du nom, etc. Ces critères, hétérogènes, mêlant des aspects lexicaux, morphologiques, syntaxiques, sémantiques voire textuels (comme le rapport type/token), sont dans ces études identifiés à priori. Pour notre part nous nous intéressons plus spécifiquement aux séquences récurrentes, car nous pensons qu'elles permettent – c'est là notre hypothèse de départ – d'intégrer les niveaux lexico-syntaxiques et discursifs, sans partir d'une typologie préalable. Nous rejoignons ainsi les travaux de Lefer, Bestgen & Grabar (2016) qui ont évalué l'apport de l'utilisation des n-grammes et de l'analyse factorielle des correspondances pour comparer des genres textuels dans plusieurs langues. Tout comme eux, nous pensons que les séquences répétées jouent un rôle pour mettre en évidence des différences discursives et rhétoriques entre les langues et entre les genres.

Selon une méthodologie inductive *corpus driven*, nous chercherons à dégager sans a priori lesquelles de ces séquences sont les plus discriminantes dans la comparaison des deux sous-genres retenus. Nous chercherons également à distinguer des traits génériques des éléments caractéristiques des auteurs, si l'on en détecte : il est possible en effet que les marques stylistiques qui distinguent un auteur parmi les autres se situent sur d'autres plans que les traits génériques proprement dits. Par ailleurs, notre

corpus étant bilingue, nous chercherons à déterminer s'il existe des traits en communs d'une langue à l'autre pour chaque sous-genre, et nous verrons si certaines caractéristiques ne sont pas transposables, pour des raisons linguistiques, culturelles ou autres.

2. Méthodologie

Nous avons réuni pour cette étude un corpus comptant deux cents œuvres, représentant, dans chaque langue, environ 5 millions de mots de romans policiers, et autant de romans de science fiction. Parallèlement, nous disposons d'un corpus de référence comptant deux cents œuvres de littérature « générale » dans chaque langue. Ce corpus nous permettra de faire émerger, par contraste, des caractéristiques communes à ces deux sous-genres. La classification des œuvres a été effectuée à priori selon des critères éditoriaux et littéraires, par des annotateurs humains.

Les sous-corpus analysés dans cette étude sont constitués de textes originaux contemporains en anglais et en français. Chaque sous-corpus compte 8 millions de mots :

- POL anglais : 47 auteurs, 69 œuvres ;
- SF anglais : 47 auteurs, 67 œuvres ;
- POL français : 46 auteurs, 69 œuvres ;
- SF français : 36 auteurs, 75 œuvres.

Pour identifier les séquences répétées, nous utilisons la méthode des arbres lexico-syntaxiques récurrents, ou ALR, qui a déjà été appliquée pour analyser les routines spécifiques aux articles scientifiques en sciences humaines et sociales (Tutin & Kraif, 2016) et les constructions lexico-syntaxiques spécifiques au roman policier et à la science-fiction (Kraif et al., 2016). Comme dans cette dernière étude, nous chercherons à faire émerger des récurrences stylistiques et thématiques. Mais le corpus sera cette fois plus étendu et comportera une dimension contrastive puisque la méthode sera également appliquée à des corpus anglais. Par ailleurs, nous chercherons à dériver, à partir des ALR, des motifs plus généraux combinant items lexicaux et syntaxiques, à l'instar des motifs d'itemsets décrits par Quiniou et al. (2012) et les motifs mixtes de Legallois et al. (2016) utilisés pour pour trouver les clichés caractéristiques du roman sentimental.

Enfin, à partir des séquences et motifs identifiés comme étant spécifiques à chacun des corpus (la spécificité étant calculée par une mesure de rapport de vraisemblance), nous utiliserons deux méthodes distinctes permettant d'identifier les groupes de séquences les plus discriminants pour la classification : comme Biber (2014) et Lefer et al. (2016), nous appliquerons des méthodes d'analyse multivariées (analyse factorielle des correspondances) afin d'identifier, par une analyse qualitative, les familles d'expressions autour desquelles s'opèrent le partage des genres. Dans un second temps, nous mettrons en œuvre une technique de classification supervisée, de type SVM ou Linear Least Squares Fit (Yang & Liu, 1999) afin de déterminer par apprentissage quels sont les traits les plus discriminants. Une évaluation de la méthode permettra de quantifier concrètement l'efficacité des critères retenus.

L'analyse qualitative des résultats nous permettra d'esquisser, au bout d'un cheminement complètement guidé par le corpus, une typologie des séquences récurrentes qui constituent selon nous la trame souvent inconsciente – ou tout au moins inaperçue – des sous-genres littéraires étudiés.

Références bibliographiques

Biber, D. (2012). Register and discourse analysis. In J. P. Gee & M. Handford (Éd.), *The Routledge handbook of discourse analysis* (p. 191-208). London ; New York : Routledge.

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast,* 14(1), 7-34. https ://doi.org/10.1075/lic.14.1.02bib

Biber, D. & Conrad, S. (2009) *Register, Genre, and Style.* New York, NY : Cambridge University Press.

Boyer, A.-M. (1992). *La paralittérature (1. éd).* Paris : Presses Univ. de France.

Kraif, O., Novakova, I., & Sorba, J. (2016). Constructions lexico-syntaxiques spécifiques dans le roman policier et la science-fiction. *Lidil. Revue de linguistique et de didactique des langues,* (53), 143-159.

Lefer, M.-A., Bestgen, Y., & Grabar, N. (2016). Vers une analyse des différences interlinguistiques entre les genres textuels : étude de cas basée sur les n-grammes et l'analyse factorielle des correspondances. Consulté à l'adresse http://natalia.grabar.perso.sfr.fr/publications/lefer-TALN2016short.pdf

Legallois, D., Charnois, T., & Poibeau, T. (2016). Repérer les clichés dans les romans sentimentaux grâce à la méthode des « motifs ». *Lidil. Revue de linguistique et de didactique des langues,* (53), 95-117.

Quiniou, S., Cellier, P., Charnois, T., & Legallois, D. (2012). Fouille de données pour la stylistique : cas des motifs séquentiels émergents. In *Journées Internationales d'Analyse Statistique des Données Textuelles (JADT'12)* (p. 821–833). Consulté à l'adresse https://hal.archives-ouvertes.fr/hal-00675586/

Siepmann, D. (2015). A corpus-based investigation into key words and key patterns in post-war fiction. *Functions of Language,* 22(3), 362-399. https://doi.org/10.1075/fol.22.3.03sie

Siepmann, D. (2016). Lexicologie et phraséologie du roman contemporain : quelques pistes pour le français et l'anglais. *Cahiers de lexicologie,* (108), 21-41. https://doi.org/10.15122/isbn.978-2-406-06281-3.p.0021

Tutin, A., & Kraif, O. (2016). Routines sémantico-rhétoriques dans l'écrit scientifique de sciences humaines : l'apport des arbres lexico-syntaxiques récurrents. *Lidil. Revue de linguistique et de didactique des langues,* (53), 119–141.

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (p. 42–49). ACM. https://doi.org/10.1145/312624.312647

Session 4.B.
Lexique spécialisé

Pour une description lexico-sémantique des verbes dans les textes spécialisés. Application multilingue aux domaines environnemental et médical

Beatriz Sánchez-Cardenas [1] et Cécile Frérot [2]
[1]Université de Grenade, Groupe de recherche Lexicon
[2]Université Grenoble Alpes, ILCEA4-GREMUTS
bsc@ugr.es, Cecile.Frerot@univ-grenoble-alpes.fr

1. Introduction

Le fonctionnement du verbe dans les textes spécialisés mérite une attention particulière car il est souvent le parent « pauvre » des dictionnaires spécialisés et des ressources terminologiques à la disposition du traducteur (L'Homme 1998 ; 2012) ; ces dernières comprennent principalement des termes à vocation nominale et si les nominalisations de verbes sont répertoriées, les verbes sont en revanche souvent omis (L'Homme (2016) donne l'exemple de *recyclage, déboisement* et *déversement* qui pourront être retenus mais pas toujours *recycler, déboiser* ou *déverser*). Par ailleurs, ce type de ressources ne met guère l'accent sur le recensement d'informations linguistiques (notamment la structure argumentale des verbes), la priorité étant donnée aux relations entre les termes et aux connaissances véhiculées par ces derniers. Les dictionnaires de langue générale quant à eux ne contiennent qu'une partie du lexique spécialisé et ne rendent pas compte des idiosyncrasies propres aux textes spécialisés ; dans le cas des dictionnaires bilingues généraux, ressources largement utilisées par les traducteurs professionnels (Josselin-Leray 2005), les entrées verbales et les informations lexico-syntaxiques qui y sont associées sont souvent lacunaires, ne reflétant pas - ou bien alors partiellement - l'usage spécialisé pourtant identifié dans des corpus (Frérot & Josselin-Leray 2008 dans le domaine de la volcanologie).

Dans la lignée des travaux visant à intégrer les arguments dans la description des termes (Lerat 2002 ; L'Homme 2012), nous cherchons à représenter la structure argumentale de verbes associés à des termes à partir de l'extraction de données provenant des textes spécialisés dans les domaines environnemental et médical. Dans la perspective appliquée d'aide à la traduction et à la rédaction spécialisée qui est la nôtre (extension de la base de données actuelles EcoLexicon au traitement de données phraséologiques), le fonctionnement linguistique des verbes en corpus mérite d'être appréhendé et recensé car pour produire ou traduire des textes qui soient fidèles aux idiosyncrasies des textes spécialisés, les rédacteurs ou les traducteurs ont besoin d'accéder à ces informations verbales spécifiques.

2. Objectifs et contexte de l'étude

La description des structures phraséologiques passe nécessairement par le référencement des unités verbales se rapportant aux termes, afin de pouvoir rendre compte des combinaisons verbo-nominales dans les discours spécialisés (Faber & López-Rodríguez 2012 ; L'Homme 2012 ; Buendía-Castro 2013 ; Bak & Novakova 2013). Par exemple, la description du terme *eaux usées* devrait être accompagnée de verbes comme *traiter, épurer, déverser, évacuer, dépolluer, décanter,* afin de permettre au traducteur ou au rédacteur d'être fidèles aux idiosyncrasies des textes spécialisés.

Dans cette perspective, nous menons une analyse linguistique à partir de corpus spécialisés dont la portée se veut multidimensionnelle en termes de langues (français-anglais-espagnol) et de domaines de spécialité (sciences de l'environnement, médecine). Les corpus, constitués de textes scientifiques dans les domaines en question en anglais, français et espagnol, comprennent chacun environ 1 million de mots. Au-delà de l'observation de données empiriques à partir de corpus, nous cherchons à explorer comment les modèles théoriques de la MGL (modèles théoriques de Grammaire-Lexicale) et de la Terminologie basée sur des cadres sémantiques (Fillmore 2006 ; León et al. 2009 ; Reimerink & Faber 2009 ; Faber 2015 ; L'Homme 2012a ; 2016) peuvent contribuer à rendre compte des propriétés linguistiques des verbes dans les textes spécialisés et aider à représenter les connaissances spécialisées des domaines. En terminologie, le modèle des cadres sémantiques a déjà porté ses fruits pour la création de ressources terminologiques comme DiCoInfo, DicoEnviro (L'Homme 2012a) et EcoLexicon (Faber et al. 2014). Son intérêt réside dans le fait qu'étant donné que les cadres sont des conceptions cognitives

universelles, les correspondances sémantiques entre les différentes langues sont plus aisées à établir lorsque les unités linguistiques sont classées dans chaque langue selon leur cadre. Cela contribue à la question de l'équivalence en traduction étant donné que l'alignement des structures argumentales dans différentes langues est un excellent moyen de prédire la traduction des verbes (Buendía-Castro & Sánchez Cárdenas 2016).

3. Analyse et premiers résultats

Dans cette étude, nous nous focalisons sur des verbes qui présentent une certaine complexité sémantique de part leur polysémie en langue générale. Nous faisons l'hypothèse que leur fonctionnement dans des textes spécialisés présente des divergences par rapport à la langue générale qu'il sera possible de mettre au jour grâce à une analyse lexico-sémantique (L'Homme et al. 2016).

Nous avons dans un premier temps observé les descriptions proposées dans la langue générale pour les verbes soumis à l'étude (verbes comme *traiter, éliminer* et *désinfecter*). Nous avons ensuite utilisé l'outil SketchEngine [1] pour analyser nos corpus et extraire les informations sur la structure argumentale de ces verbes dans les domaines médical et environnemental. Nous présentons ici notre analyse pour le verbe *traiter*.

Selon les informations contenues dans le Trésor de la Langue Française informatisé (TFLi), les usages du verbe *traiter* appartiennent à quatre domaines lexicaux différents :

1. Traiter qqn d'imbécile : **COGNITION** | **EMOTION**
   Mon chef m'a traité d'inutile.

2. Traiter quelque chose : **ACTION**
   Ce système informatique traite les transactions enregistrées par les systèmes de paiement électronique.

3. Traiter d'un sujet : **SPEECH**
   Angela Merkel traite avec François Hollande la sortie de l'euro de la Grèce.

4. Traiter avec un produit : **MANIPULATION**
   Nous avons traité le meuble avec de la cire.

L'observation des données issues de nos corpus montre que ces usages deviennent restreints dans les domaines de spécialité étudiés. Les arguments de ces verbes étant différents en fonction du domaine de spécialité (médical et environnemental), ils donnent lieu à des configurations sémantiques et à des structures argumentales propres à chaque domaine. Le tableau 1 montre les différences de restriction lexicale que le verbe *traiter* impose à ses arguments.

Langue générale		Domaine médical		Domaine environnemental	
Agent	Thème	Agent	Thème	Agent	Thème
sujet	demande	médecin	hypertension	station d'épuration	effluent
dossier	problème	médicament	symptôme	installation	déchet
transaction	question	thérapie	infection		eaux d'égouts
thématique	affaire	médicament	maladie		eaux usées
livre	question		acné		graisses
document	demande		cancer		nitrate
			(...)		(...)

Table 1 – Arguments du verbe *traiter*

Ces résultats nous orientent vers la configuration du cadre sémantique du verbe *traiter* dans chaque domaine. Prenons comme exemple le concept du *"traitement des eaux usées"*. Notre étude de corpus montre que le seul domaine lexical activé est celui de **MANIPULATION** :

1. Accessible en ligne (version démo) à l'adresse : www.sketchengine.co.uk

« Une station d'épuration d'eaux manipule les eaux usées pour enlever les polluants afin de les rendre propres. »

Du point de vue de la sémantique des cadres, le verbe *traiter* a deux actants dans ce domaine de spécialité : une entité A qui intervient activement pour transformer une entité B de façon positive. Ainsi, une situation au départ négative devient au final une situation positive. Le verbe rentre ici dans le cadre sémantique « processing_materials ». La définition du cadre sémantique se référant à toutes les langues décrites (anglais, français, espagnol), il apparait par convention en anglais :

An Agent alters some Material in some useful way by means of some chemical or physical Alterant. Typically, this involves placing a reagent in contact with the Material, or applying heat, pressure, etc. ¶

Le référent de l'Agent est souvent un objet construit par l'homme, qui est en réalité le responsable « caché » du changement :

Cette station (Agent) peut traiter les eaux (Material) d'une ville d'un million d'habitants environ. ¶
Les installations (Agent) traitent l'ensemble des eaux usées (Material). ¶

À partir de ces informations, il est possible de construire le cadre sémantique multilingue du verbe *traiter* (Tableau). En effet, l'étude de cette structure argumentale dans les corpus anglais et espagnol montre que les verbes qui figurent dans cette configuration lexico-grammaticale sont les verbes *treat* et *depurar*.

	Domaine lexical : ACTION Cadre sémantique : Processing_materials		
	Argument 1	Verbe	Argument 2
rôle sémantique	Agent	TREAT (en) TRAITER (fr) DEPURAR (es)	Material
classe lexicale	ARTEFACT		MATERIEL ORGANIQUE
Exemple	*La station d'épuration des eaux usées de Doubs*		***les eaux usées** domestiques et industrielles de 24 communes.*

TABLE 2 – Structure argumentale multilingue de trois verbes équivalents

Nous observons que l'usage du verbe en langue de spécialité implique une restriction aussi bien du sens que des structures syntaxiques. Cela tend à montrer, d'une part que les structures syntaxiques sont porteuses de sens et, d'autre part, qu'un verbe de la langue générale devient spécialisé en langue de spécialité en raison des termes qu'il sélectionne comme argument. C'est ce comportement verbal que nous souhaitons appréhender pour l'intégrer dans notre base de données phraséologiques et ainsi contribuer à la mise à disposition d'informations linguistiques utiles aux utilisateurs (traducteurs ou rédacteurs). Dans notre communication, nous nous attacherons à montrer l'intérêt de ces verbes dits « généraux » pour la description des langues de spécialité. Nous exposerons par ailleurs la démarche suivie pour extraire du corpus les informations nécessaires à la construction d'un cadre sémantique en langue de spécialité avant de présenter un modèle de description de l'équivalence verbale en langue de spécialité.

Références bibliographiques

Bak, M., & Novakova, I. (2013) Le raisonnement dans les textes scientifiques : le cas des verbes causatifs. In A. Tutin & F. Grossmann (Éd.), *L'Écrit Scientifique : Du Lexique Au Discours. Autour De Scientext.* Rennes : Presses Universitaires de Rennes, 101-122.

Buendía Castro, M. (2013) *Phraseology in Specialized Language and its Representation in Environmental Knowledge Resources.* PhD Thesis. Universidad de Granada, Granada, Spain.

Buendía-Castro, M. & Sánchez-Cárdenas, B. (2016) Using Argument Structure to Disambiguate Verb Meaning. In *Proceedings of the XVII EURALEX international congress,* edited by Margalitadze, T. & Meladze, G., pages 482-490. Tbilisi : Ivane Javakhishvili Tbilisi University Press.

Faber, P, & López Rodríguez, C.-I. (2012). Terminology and Specialized Language. In Faber P. (éd) *A Cognitive Linguistics View of Terminology and Specialized Language.* Berlin, Boston : De Gruyter Mouton. 9-31.

Faber, P., León Araúz, P. & Reimerink, A. (2014) Representing environmental knowledge in EcoLexicon. In *Languages for Specific Purposes in the Digital Era.* Educational Linguistics, 19 :267-301. Springer.

Faber, P. (2015) Frames as a framework for terminology. In Kockaert, H.J. & Steurs, F. (éds) *Handbook of Terminology,* 1 :14-33. John Benjamins Publishing Company.

Fillmore, C. J. (2006). Frame semantics. In D. Geeraerts (Ed.), *Cognitive linguistics : Basic readings.* Berlin : Mouton de Gruyter, 373-397.

Frérot, C. Josselin-Leray, A. (2008). Contribution des corpus à l'enrichissement des dictionnaires bilingues généraux. Application au domaine de la volcanologie. *Autour des langues et du langage. Perspective interdisciplinaire.* Presses Universitaires de Grenoble, 415-422.

Josselin-Leray, A. (2005). *Place et rôle des terminologies dans les dictionnaires généraux unilingues et bilingues. Etude d'un domaine de spécialité : volcanologie,* thèse de doctorat, Université Lyon II.

L'Homme, M.-C. (2012). Le verbe terminologique : un portrait de travaux récents. In Neveu, F. et al. (éd).*Actes du 3e Congrès mondial de linguistique française,* Lyon, France, EDP Sciences.

L'Homme, M.-C. (2012a). Adding syntactico-semantic information to specialized dictionaries : an application of the FrameNet methodology. In Gouws, R. et al. (éds.). *Lexicographica 28,* 233-252.

L'Homme, M.-C. (2016). Terminologie de l'environnement et Sémantique des cadres. In *SHS Web of Conferences* (Vol. 27, p. 05010). EDP Sciences.

L'Homme, M.-C. (1998). Le statut du verbe en langue de spécialité et sa description lexicographique. *Cahiers de lexicologie,* 73(2), 61–84.

L'Homme, M.-C., Subirats, C., & Robichaud, B. (2016). A Proposal for combining "general" and specialized frames. *COLING 2016,* 156.

León Araúz, P., Reimerink, A. & Faber, P. (2009). Puertoterm & Marcocosta : A Frame-based Knowledge Base for the Environmental Domain. *Journal of Multicultural Communication, 1* :47-70.

Lerat, P. (2002). Qu'est-ce que le verbe spécialisé ? Le cas du droit. *Cahiers de Lexicologie, 80*, 201–211.

Reimerink, A. & Faber, P. (2009). Ecolexicon : A frame-based knowledge base for the environment. In *European conference of the Czech Presidency of the Council of the EU TOWARDS eENVIRONMENT Opportunities of SEIS and SISE : Integrating Environmental Knowledge in Europe,* edited by Hrebícek, J., Mírovský, J.H.a.P.a., Pillmann, W., Holoubek, I. & Bandholtz, T., pages 25-27. Brno : Masaryk University

Elaboration d'un lexique scientifique trans-biomédical

Anastasia Galmiche et Izabella Thomas
Centre de Recherches Interdisciplinaires et Transculturelles (C.R.I.T.), Université Bourgogne Franche-Comté
anastasia.galmiche@edu.u-fcomte.fr, izabella.thomas@univ-fcomte.fr

1. Contexte et objectifs

L'attractivité et la qualité d'une recherche universitaire sont étroitement liées à la diffusion internationale de ses résultats. Aujourd'hui, la plupart des publications scientifiques se font en anglais, surtout en ce qui concerne les sciences dites 'dures', dont le biomédical fait partie. Cependant, la barrière linguistique constitue un obstacle important devant la publication en langue anglaise pour de nombreux chercheurs francophones. Des travaux de recherche ont été consacrés à la spécificité des textes scientifiques, modelés sur des standards anglo-saxons, et des manuels de rédaction scientifique sont proposés aux utilisateurs. Les outils informatisés sont pour la plupart destinés aux traducteurs et consistent en des ressources généralistes, non-centrées sur la traduction scientifique : traducteurs automatiques, mémoires de traduction, bases de données terminologiques, les dictionnaires/glossaires électroniques mono/multilingues. Il existe aussi des outils informatisés plus spécifiques, mais ils sont totalement inconnus du public biomédical rédigeant.

Notre objectif consiste à concevoir un logiciel d'aide à la rédaction scientifique dans le domaine biomédical. Pour ce faire, nous avons d'abord défini le périmètre d'un tel logiciel, en nous appuyant sur l'état de l'art et sur une enquête auprès des acteurs du terrain, les experts en rédaction médicale ainsi que les médecins, chercheurs et traducteurs dans le biomédical [1]. Il en résulte plusieurs orientations méthodologiques, dont la première consiste à élaborer ce que nous appelons le lexique scientifique trans-biomédical.

Nous définissons le *lexique scientifique trans-biomédical* comme un ensemble du lexique médical et des collocations autour de ce lexique, présents dans plusieurs sous-domaines du biomédical sous ses formes utilisées/utilisables dans des documents scientifiques. Cela veut dire que l'on peut retrouver ce lexique dans n'importe quel sous-domaine du biomédical, puisqu'il n'est pas spécifique à un sous-domaine en particulier (exemples : *health, patient, syndrome, disease...*)

Pour définir le lexique trans-biomédical nous nous appuyons sur le concept de *lexique scientifique transdisciplinaire* qui est constitué des lexies servant à décrire les activités scientifiques et la méthodologie de la recherche à travers les différentes disciplines.

2. État de l'art

Les premières listes de vocabulaire transdisciplinaires ont été créées dans les années 1970 : il s'agissait de l'Academic Vocabulary List (CAMPION et ELLEY 1971) et l'American University Word List (PRANINSKAS 1972). Les chercheurs se sont appuyés sur les critères de fréquence et de répartition des mots dans un corpus multidisciplinaire pour inclure des lexies dans ces listes. GHADESSY et LYNN (GHADESSY 1979 ; LYNN 1973) ont choisi de regrouper dans leurs listes les mots les plus fréquemment annotés par les étudiants dans les manuels, une annotation représentant la difficulté à retenir un mot considéré comme important pour un étudiant. Ces quatre listes ont ensuite été regroupées dans la University Word List (XUE et NATION 1984) qui au total comprend 800 mots représentatifs.

La liste la plus connue est l'Academic Word List, établie dans les années 2000 par COXHEAD (2000) à l'aide d'un corpus de 3 500 000 mots provenant de manuels universitaires portant sur 28 disciplines différentes. Cette liste est constituée des 570 mots représentatifs de familles de mots ayant une fréquence minimum de 100 et qui sont utilisés dans au moins 14 domaines différents. Ces 570 familles de mots couvrent 10% des mots du corpus.

Au fil des années, ont émergé des listes académiques basées sur une discipline unique, telles que l'informatique (LAM 2001), les affaires (HSU et al. 2011), l'ingénierie (MUDRAYA 2006), l'agriculture (MARTINEZ, BECK et PANZA 2009), le journalisme (CHUNG 2009) ou bien la théologie (LESSARD-CLOUSTON 2006).

1. Une enquête destinée aux professionnels de la Santé, diffusée en ligne entre 01/12/2015 et 31/01/2016.

Pour la médecine, la Medical Academic Word List (MAWL) a été proposée par WANG et al. (2008). Cette liste réunit 623 mots communs aux sous-disciplines biomédicales les plus représentées dans les écrits biomédicaux (Tableau 1). Elle est construite à partir d'un corpus composé de 288 textes provenant de 32 disciplines médicales et écrits par au moins un auteur anglophone. Le corpus comprend au total 1 093 011 mots appartenant à 31 275 familles de mots différentes.

La méthodologie s'appuie sur les critères utilisés par COXHEAD (2000) pour la création de l'Academic Word List, en y ajoutant le critère de spécificité. Ainsi, pour qu'une famille de mots soit retenue dans la MAWL, elle doit répondre aux critères suivants :

- elle doit être présente dans au moins 16 sous-domaines médicaux ;
- elle doit apparaître plus de 30 fois dans le corpus ;
- elle doit être spécifique au domaine médical.

Afin de vérifier le dernier critère, WANG a fait appel à des experts du domaine biomédical. De manière assez surprenante, ils ont décidé de supprimer 27 familles de mots car jugés trop spécialisées (par exemple : pathogenesis, cytokine, epithelial etc.).

De l'autre coté la liste contient des mots tels que *whereas* and *thereby*, qui ne sont, de toute évidence, pas spécifiques du domaine biomédical. Le problème de la liste de WANG réside dans le fait que l'appartenance d'un mot à la liste est décidée sur le mot lui-même sans avoir recours au contexte. Or, comme le constate FRASER (2007, 2009a, 2009b), la majorité des mots retenus selon les critères de fréquence et de répartition (donc, les critères de WANG) sont ambigus du point de vue de leur spécialisation dans le domaine. Il existe des mots qui ont pour origine un sens spécialisé mais qui sont communément utilisés dans le lexique général ; d'un autre côté, les termes crypto-techniques sont des mots du lexique général ou transdisciplinaire qui acquièrent un sens spécialisé dans un domaine. Fraser décide donc de procéder à une étude des cooccurrents afin de déterminer le sens véhiculé par l'unité lexicale analysée.

L'étude des cooccurrents est nécessaire pour la détermination du sens mobilisé par une unité lexicale. En effet, si une unité lexicale a pour cooccurrents des termes du domaine étudié, alors il est fort probable qu'elle représente elle-même un terme de ce domaine. Ainsi, Fraser a déterminé que le verbe *block* était utilisé dans le sens « empêcher l'action d'une drogue » car les collocations fréquentes dans lequel block apparait sont *blockade of [receptor], channel blocker(s), beta blocker(s).*

3. Méthodes et résultats

Compte tenu de nombreuses critiques concernant les résultats de Wang, nous avons tout d'abord décidé de vérifier la pertinence du lexique retenu par Wang sur un corpus plus grand, à savoir le corpus PLOS [2]. C'est un corpus de 46000 articles extraits de 5 revues PLOS différentes : PLOS Biology, PLOS Computational Biology, PLOS Medicine, PLOS Pathogens et PLOS Neglected Tropical Diseases. L'expérimentation de Wang a été adaptée au corpus PLOS, qui contient 41 millions de mots et que nous avons divisé en 46 sous-domaines. Par conséquent, les critères ont été réajustés : pour la répartition, un mot doit appartenir à au moins 23 domaines et pour la fréquence, avoir une fréquence d'au moins 30 occurrences pour 1 millions de mots (donc 1230 occurrences dans l'ensemble du corpus).

Sur 623 familles de mots de la liste de WANG, 53 familles (8%) n'ont pas respecté ces critères. Le critère non tenu est plutôt celui de la fréquence (53 familles) que de la répartition (3 familles : *ration, perception et append*). Parmi les 53 familles qui ne respectent pas le critère de la fréquence, la majorité n'est pas spécifique au biomédical : *concomitant, aknowledge, comment, inferior, thereafter*, etc., mais certaines correspondent bien à des lexies spécialisées : *methanol, catheter, laser*, etc. En plus, la répartition de mots selon la fréquence ne correspond pas dans les deux expérimentations : on ne retrouve pas les termes les plus et les moins fréquents aux mêmes rangs dans les deux listes. Par exemple, si on fait la comparaison entre les 50 familles de mots les moins fréquentes dans les deux

2. La *Public Library of Science*, Bibliothèque scientifique publique, un projet de publication scientifique anglophone à accès ouvert fonctionnant sur la base de licences libres ; nous remercions l'OST, Montréal pour nous avoir donné le droit d'utiliser ce corpus.

corpus, on retrouve seulement 19 familles en commun. Dans les résultats les plus fréquents, on retrouve 27 familles en commun dans les premiers 50 familles de chaque liste. En général, plus on va vers les résultats fréquents, plus les mots correspondent à des lexies spécialisés (*cell, gene, infect*, etc.).

4. Conclusion

Cette évaluation démontre que le lexique médical tel que proposé par Wang n'est pas stable en fonction du corpus utilisé. De plus, si nous voulons prendre appui sur la MAWL pour le lexique trans-biomédical, il faut non seulement supprimer les lexies spécialisées qui n'ont pas une fréquence suffisante dans un corpus plus large, mais aussi procéder au tri entre lexies spécialisées et non-spécialisées. Ceci nous permet, dans un premier temps, d'établir 320 familles de mots potentiellement retenues pour le lexique trans-biomédical (Tableau 2). Cependant, avant de proposer une liste définitive, nous souhaitons élargir notre expérimentation dans deux directions. Premièrement, nous allons comparer notre liste à une nouvelle proposition de Medical Academic Vocabulary List, établie en 2016 par Lei et Liu (2016), qui, eux aussi, critiquent l'approche de Wang et introduisent de nouveaux critères concernant la discrimination entre le vocabulaire spécialisé et le vocabulaire général. Deuxièmement, nous souhaitons vérifier la pertinence sur le corpus du lexique provenant d'un dictionnaire biomédical spécialisé. Pour cela, nous allons vérifier les critères de fréquence et de distribution des termes provenant du Dorland's Illustrated Medical Dictionary (2012), un dictionnaire biomédical contenant plus de 120 000 termes biomédicaux, régulièrement mis à jour.

Numéro	Mot représentatif	Numéro	Mot représentatif	Numéro	Mot représentatif
1	cell	11	tissue	21	therapy
2	data	12	dose	22	indicate
3	muscular	13	gene	23	area
4	significant	14	previous	24	obtain
5	clinic	15	demonstrate	25	research
6	analyze	16	normal	26	vary
7	respond	17	process	27	activate
8	factor	18	similar	28	require
9	method	19	concentrate	29	induce
10	protein	20	function	30	cancer

TABLE 1 – **Premiers mots de la MAWL (WANG et al. 2008), rangés par fréquence**

Numéro	Mot représentatif	Numéro	Mot représentatif	Numéro	Mot représentatif
1	cell	11	formation	21	process
2	gene	12	clinic	22	score
3	infect	13	induce	23	drug
4	analyze	14	function	24	area
5	protein	15	region	25	tissue
6	indicate	16	site	26	secondary
7	differentiate	17	select	27	range
8	sequence	18	strain	28	cluster
9	factor	19	inhibit	29	structure
10	virus	20	positive	30	bacterium

TABLE 2 – **Premiers mots du lexique trans-biomédical, rangés par fréquence**

Références bibliographiques

Campion, M. E., & Elley, W. B. (1971). *An academic vocabulary list.* Wellington, N.Z. : New Zealand Council for Educational Research.

Chung, M. (2009). The newspaper word list : A specialised vocabulary for reading newspapers. *JALT Journal, 31(2)*, 159–182.

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly, 34(2)*, 213–238.

Dorland, W. A. N. (2012). *Dorland's illustrated medical dictionary.* Philadelphia, PA : Saunders/Elsevier.

Fraser, S. (2007). Providing ESP Learners with the Vocabulary They Need : Corpora and the Creation of Specialized Word Lists. *Hiroshima Studies in Language and Language Education, Issue 12*, 127–143.

Fraser, S. (2009a). Breaking Down the Divisions between General, Academic, and Technical Vocabulary : The Establishment of a Single, Discipline-based Word List for ESP Learners. *Hiroshima Studies in Language and Language Education, (12)*, 151–167.

Fraser, S. (2009b). Technical vocabulary and collocational behaviour in a specialised corpus. *Proceedings of the British Association for Applied Linguistics (BAAL)*, 3-5.

Ghadessy, P. (1979). Frequency counts, word lists, material preparation : A new approach. *English Teachung Forum*, (17).

Hsu, W., & al. (2011). A business word list for prospective EFL business postgraduates. *The Asian ESP Journal, 7(4)*, 63–99.

Lam, J. K. M. (2001). A study of semi-technical vocabulary in computer science texts, with special reference to ESP teaching and lexicography. *In G. James (Ed.)* Research Reports.

Lei, L., & Liu, D. (2016). A new medical academic word list : A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes*, 22, 42–53.

Lynn, R. W. (1973). Preparing Word-Lists : A Suggested Method. *RELC Journal.*

Lessard-Clouston, M. (2006). Breadth and depth specialized vocabulary learning in theology among native and non-native English speakers. *Canadian Modern Language Review, 63(2)*, 175–198.

Martinez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles : A corpus-based study. *English for Specific Purposes, 28(3)*, 183–198.

Mudraya, O. (2006). Engineering English : A lexical frequency instructional model. *English for Specific Purposes, 25(2)*, 235–256.

Praninskas, J. (1972). *American university word list.* Longman Group Limited.

Wang, J., Liang, S., & Ge, G. (2008). Establishment of a Medical Academic Word List. *English for Specific Purposes, 27(4)*, 442–458.

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication, 3(2)*, 215–229.

Session 5.A.
Dialogisme et philosophie

Les emplois en « c'est » dans le corpus Philosophèmes : définition ou exemplification ?

Emmanuèle Auriac-Slusarczyk [1], Mylène Blasco [2] et Philippe Roiné [3]
[1] Université Clermont Auvergne, ACTé, F-63000 Clermont-Ferrand, France
[2] Université Clermont Auvergne, LRL, F-63000 Clermont-Ferrand, France
[3] Université Cergy-Pontoise, EMA, F-F-95000 Cergy-Pontoise, France
emmanuele.auriac@uca.fr, mylene.blasco-bulbecco@uca.fr, philippe.roine@u-cergy.fr

1. Introduction

Depuis 10 ans, une collaboration scientifique pluridisciplinaire grandit autour d'un corpus de discussions à visée philosophique (Auriac-Slusarczyk, E. & Blasco-Dulbecco, M. 2013 ; voir http://philosophemes.univ-bpclermont.fr/). Ce corpus fait désormais l'objet de présentations récurrentes dans divers colloques, à visée de recherche fondamentale en sciences du langage (Blasco & Auriac-Slusarczyk, 2016) ou à visée éducative (Auriac-Slusarczyk & Slusarczyk, 2015).

1.1. Le Corpus Philosophèmes : présentation

Les ateliers philosophiques se pratiquent depuis une trentaine d'années dans plusieurs pays. Le philosophe Matthew Lipman (1995) a transposé l'investigation de type philosophique depuis l'enseignement supérieur jusqu'aux classes des écoles maternelles en faisant réfléchir des élèves dans des communautés de recherche à visée philosophique (Auriac-Slusarczyk & Blasco-Dulbecco, 2013 ; Auriac-Slusarczyk & Coletta, 2015 ; Simon & Tozzi, à paraître 2017). Les enseignants qui ont participé au recueil des données ont appliqué les principes de cet ajustement. Le corpus philosophèmes, 30 discussions d'environ 40 minutes, est constitué de paroles d'élèves, enfants (de 6 à 12 ans) et adolescents (de 14 à 18 ans) enregistrées au sein de ces ateliers philosophiques lors de discussions tenues en classe à l'école primaire et au collège entre 2010 et 2013 et transcrites selon des conventions qui respectent totalement la langue dans ce qu'elle produit en spontané. Ces données verbales et co-verbales forment un corpus qui a pour caractéristiques d'être transversal et longitudinal. Il est stocké pour permettre une visée d'étude interactionniste de la langue (http://philosophemes.univ-bpclermont.fr/). Les données sont fiables, non corrigées, non falsifiées. Le travail de constitution de ce corpus et son analyse a bénéficié de soutiens financiers dans le cadre de 5 projets ou opérations de recherche successifs (Daniel, 2009 ; Specogna et Halté, 2009 ; Auriac-Slusarczyk & Blasco-Dulbecco, 2010 ; Auriac-Slusarczyk & Lebas-Fraczak, 2011 ; Specogna & Saint-Dizier de Almeida, 2012).

Destiné au partage scientifique, ce corpus a donné matière à des travaux nombreux et variés dans des champs disciplinaires différents (voir Auriac-Slusarczyk & Colletta, 2015). Le corpus s'est récemment agrandi en accueillant les données issues du travail doctoral de Philippe Roiné (Roiné, 2016), qui concerne exclusivement le niveau d'élèves de CM1 /CM2. Les élèves dont les paroles en ateliers ont été transcrites selon les normes du corpus *Philosophèmes* sont orientés sur une activité définitionnelle. Il comporte des emplois en « c'est» dont le fonctionnement valide et complète les emplois déjà observés dans l'ensemble du corpus *Philosophèmes*.

1.2. Le corpus Philosophèmes : orientation vers une exploration écologique

Notre approche est écologique, au double regard de la spécificité des données et de l'objet d'investigation linguistique déterminé a posteriori.

Pertinence des données. : Le corpus **Philosophèmes** aborde l'usage de la langue spontanée dans un contexte scolaire sur des sujets sensibles : l'amour, la mort, le mensonge, le partage, l'argent, la sécurité, les origines, l'obéissance, le courage, l'amitié. C'est un corpus nouveau sur la langue orale qui permet de disposer de paroles véritables d'élèves et d'enseignants du primaire au collège pour renouveler le regard porté sur l'activité réflexive, répertorier et illustrer la dynamique des raisonnements et les emplois verbaux associés chez les élèves. Dans les ateliers philosophiques deux facultés humaines et fondamentales sont mobilisées en même temps : celle de parler et celle de penser individuellement et collectivement, dans un même espace, sur un même sujet (les choses de la vie), entre élèves et avec

l'enseignant. L'échange collectif crée alors la nécessité régulière de définir ce dont il est question : mots, idées, concepts.

Exploration linguistique : dans l'espace de paroles sensées, les dires et les raisonnements « aboutissent ou non, s'entrechoquent, s'entremêlent, s'agrègent, dans une dynamique positive » (Auriac-Slusarczyk, Fiema et alii, 2013). La parole favorisée retranscrite admet une épaisseur intéressante. L'intérêt est de rapprocher ce qui relève de la langue, de la parole, du raisonnement, de la pensée. On a fait l'hypothèse que les élèves orientés sur des conduites de raisonnement (avec activation d'une pensée) individuelles et collectives produisent des faits de langue particuliers en partie exploités (Blasco-Dulbecco & Auriac-Slusarczyk, 2016). On peut pratiquer des fouilles *a priori* ou bien dérouler des études hiérarchiques et fonctionnelles pour détecter le tour particulier de certaines formes morphosyntaxiques : *il y a, c'est*, etc. et montrer que ces tours seraient des marqueurs de choix dans l'opération de définition.
Les points abordés concernent les structures syntaxiques de type :

1. ben pour moi l'amour **c'est** un sentiment *qui se* passe dans la tête et *que* l'on ne contrôle pas et que l'on a envers quelqu'un ou quelque chose euh plein de choses (Amour TP 78)

2. la jalousie justement en général **c'est** (...) en vrai **c'est pas** de l'amour (Amour TP 351)

Ces structures en c'est interviennent dans des formulations que l'on désignera plutôt comme étant de type explicatives, leterme de définition renvoyant davantage en linguistique à des pratiques plus codifiées (Gréco & Traverso, 2016).

2. Le travail engagé

L'étude que nous présentons a consisté à observer les faits de langue pour comprendre la concomitance entre le dire et le penser, pour conduire une double étude syntaxique et pragmatique. Nous avons investi une étude linguistique instrumentée avec Antconc pour vérifier la variété des co-textes, sélectionné les épisodes représentatifs, puis appliqué aux épisodes remarquables une grille d'étude syntaxique, enfin interprété à partir de cette fouille et de ces grilles, une étude pragmatique des usages en « c'est », et en « c'est pas ». Nous avons traité le corpus en deux phases : avant et après extension aux cinq discussions du corpus Roiné et comparé nos données à d'autres corpus. Ce que le logiciel ne peut pas voir... : apport de la mise en grille.

La mise en grille syntaxique visualise spatialement le discours oral rendant compte des modes de déploiement syntagmatique et paradigmatique du discours. La présentation en grille rompt avec la linéarité telle qu'elle apparaît dans les transcriptions « textuelles» habituelles en ajoutant une dimension verticale. Elle montre davantage les régularités et les variations qui structurent l'élaboration du discours (Blasco 2016). Elle dégage une visualisation experte sur laquelle l'interprétation pragmatique prend appui ; cette dernière valide ou transforme l'analyse syntaxique proposée. La mise en grille d'épisodes met en valeur la nature fortement interactive des discussions, en retraçant les relations syntaxiques, morphosyntaxiques, lexicales et sémantiques entre les propos de différents élèves. Une représentation en grille d'un des énoncés introduits ci-dessus illustre la progression syntaxique et sémantique telle qu'elle s'opère sur les deux axes :

3. ben pour moi l'amour c'est un sentiment qui se passe dans la tête
   et que l'on ne contrôle pas
   et que l'on a envers quelqu'un
   ou quelque chose euh
   plein de choses

2.1. Étude pragmatique et interlocutoire des propos

Etayée par l'analyse instrumentée puis par la présentation en grille de divers épisodes caractérisant des emplois contrastés de « c'est », l'interprétation pragmatique complète la description syntaxique

en intégrant un empan plus large que les co-occurrences directes. On met en exergue différentes valeurs définitionnelles des emplois en « c'est » en mettant en relation l'amont et l'aval du discours au sein d'évènements interlocutoires (Trognon, 1995, 1999) reprenant tout ou partie des épisodes fléchés par l'analyse syntaxique. La portée, la valence définitionnelle ou illustrative des emplois en « c'est » sont ainsi davantage situées dans la dynamique générale de la discussion ; la comparaison d'emplois plus ou moins caractéristiques (par confirmation/infirmation avec le corpus Roiné) dégage une manière de définir, pour partie, le genre philosophique à partir de ce corpus spécifique. C'est une nouvelle indication de l'importance que présentent les genres de discours pour l'analyse syntaxique et pragmatique (Biber & Conrad 2010). Dans ce travail, l'élargissement des données permet de repérer de nouveaux environnements, de développer de nouvelles analyses et d'examiner l'incidence de divers facteurs sur les structures en *c'est* attestées dans ces productions pour l'activité définitionnelle (Auger, 1997 ; Gréco & Traverso 2016). Pour cette question, le corpus *Philosophèmes* a été confronté à d'autres corpus (dont MPF, Multicultural Paris French) dans Blasco & Cappeau (2017 à par.)

3. Résultats

Plusieurs perspectives sont présentées : les tendances distributionnelles, cooccurrences morpho-syntaxiques, les spécificités lexicales et grammaticales, et premières concordances seront brossées. Ensuite on vérifie en quoi « c'est » est mis au service de la construction du texte dans l'espace de construction de la pensée. On montre en quoi la grammaire de la langue répond à des tendances situées non attendues qui participent à la construction inédite de raisonnements tout aussi inédits. Ensuite, les séquences définitionnelles ou exemplificatoires, celles qui comportent « c'est », sont détaillées.

3.1. Exploration instrumentée du corpus Auriac-Slusarczyk 2009/2013

Nous dépassons les conditions des études de discussions scolaires préliminaires conduites (Blasco-Dulbecco & Auriac, 2010) limités à des comparaisons en pourcentage. Nous décrirons à partir de l'exploration automatisée avec *Antconc*, l'usage de la grammaire pour révéler en quoi elle articule parole et pensée de manière originale. L'ensemble de faits observés qui ressort : 1 est décrit au plan syntaxique : cinq grilles syntaxiques sont exploitées pour illlustrer, ce sont celles qui différencient les procédés grammaticaux ; 2. les faits sont réinterprétés au plan pragmatique : reprise des grilles syntaxiques pour étudier le déploiement interlocutoire particulier à l'éclairage syntaxique. On distingue à cette étape différents emplois définitionnels/explicatifs et exemplificatoires qui ne jouent pas le même rôle dans le déploiement interlocutoire des raisonnements produits.

3.2. Fouille complémentaire croisée dans le Corpus Roiné 2012/2016

Nous recensons l'ensemble des occurrences du corpus Roiné en « c'est », de manière automatisée avec Antconc, puis sélectionnons les occurrences qui valident ou étendent les emplois déjà considérérés et décrits auparavant au plan syntaxique (cf.3.1.). Après mise en exergue de ces emplois situés extraits, répertoriés, et doublement interprétés, une fouille complémentaire dans le corpus Roiné oriente vers la validation ou l'extension des formes d'emploi en « c'est ». Trois trois approches de la définition par équivalence (ex. 4), par caractérisation (ex. 5), par désignation (ex. 6) apparaissent caractéristqiues (voir ci-dessous).

4. un animal intelligent c'est un animal qui sait apprendre (explicitation par équivalence)

5. l'amour pour une personne c'est quand on est attaché (caractérisation contextuelle)

6. donc toi pour toi venir en philo c'est du courage (explicitation par désignation)

4. Discussion

L'extraction de phénomènes morpho-syntaxiques remarquables et l'analyse pragmatique menées prouvent en quoi les ateliers de philosophie présentent des espaces inédits expliquant les modalités particulières d'inscription de l'activité de pensée dans les structures morphologiques de la langue, et

vice-versa. On discute ces aspects comparativement, d'une part, en regard des usages plus anciens de « c'est » et, d'autre part, au regard de l'intérêt que ces usages contemporains de « c'est » en ateliers de philosophie apportent face aux usages colectés dans d'autres corpus (conversations adolescentes, inteview d'aultes, etc.).

Références bibliographiques

Auger Alain. 1997. Repérage des énoncés d'intérêt définitoire dans les bases de données textuelles. Thèse. Neuchâtel. https://doc.rero.ch/record/473/files/these_AugerA.pdf.

Auriac-Slusarczyk, E. & Colletta, J-M. (2015) Les ateliers de philosophie. Une pensée collective en acte. Clermont Ferrand : Université Blaise Pascal.

Auriac-Slusarczyk, E. & Slusarczyk, B. (2015, coord.). Les ateliers de philosophie au service de la raison citoyenne. Symposium In Colloque ESREA Continuity and Discontinuity in Learning Careers : Potentials for a Learning Space in a Changing World, Spain : University of Séville 25-27 novembre .

Blasco, M. & Auriac-Slusarczyk, E, (2016). Une langue à l'instant t : les ateliers philosophiques à l'école. Colloque international, Changements linguistiques et phénomènes sociétaux, France : Lyon.

Auriac-Slusarczyk, E. & Fiema G., (2013). Raisonner et discuter : définitions et principe d'étude pragmatique du corpus philosophèmes, Cahier du LRL 5, 41-74.

Auriac-Slusarczyk, E., Lebas-Fraczak, L., Blasco, M./Daniel, M-F, Colletta J-M, Simon, J-P., Fiema, G., Auriel A. & Henrion, J. (2012). Philosophèmes, Congrès national du réseau des MSH, Quelles sciences humaines et sociales pour le 21e siècle ? Caen, 6 et 7 décembre 2012. Lien internet http://www.msh-reseau.fr/IMG/pdf/poster_philosophemes.pdf

Bautier, E. & Rayou, P. (2009). Les inégalités d'apprentissage. Paris : PUF.

Benninger, C. (2014). La question de la définition sémantique du nom atypique chose, Travaux de Linguistique 69, 35-55.

Benninger, C. (2016). ‘ Une chose [X] : P ‘ : une conjonction de contraintes. Communication au Colloque international : Contraintes linguistiques, linguistique contrainte. A propos de la complémentation nominale. Université Paris-Descartes 2-3 juin 2016.

Berrendonner, A. (2002). Les deux syntaxes, Verbum 1-2, 24, 23-35.

Biber, D. & Conrad, S. (2010). Register, Genre, and Style. Cambridge : Cambridge University Press.

Blanche-Benveniste, C. (1997). Approches de la langue parlée en français. Paris : Ophrys.

Blanche-Benveniste, C. (1991). Analyses grammaticales dans l'étude de la langue parlée, in U. Dausendschön-Gay & E. Gülich Krafft (éds). Linguistische Interaktionsanalysen (Linguistische Arbeiten 254). Tübingen : Niemeyer, 1-18.

Blasco, M. & Lebas-Fraczak, L. (2017) (à paraître). Les grilles syntaxiques comme aide à l'étude des opérations intellectuelles mises en œuvre dans les DVP, in : J.-P Simon & M. Tozzi (dir.). Paroles de philosophes en herbe. Regards croisés de chercheurs sur une discussion sur la justice en CM2. Grenoble : Ellug coll. Langues, gestes, paroles, 188-195.

Blasco, M. (2016). Une lecture grammaticale de séquences choisies dans les échanges philosophiques, in V. Saint-Dizier de Almeida & E. Auriac-Slusarczyk (dir.). Les ateliers-philo en contexte scolaire. Recherches En Education 24, 110-121.

Blasco, M. & Cappeau, P. (à par. 2017). Analyse syntaxique et contextuelle des structures SN1 c'est SN2 de type : l'amour c'est un sentiment qui se passe dans la tête

Actes du Colloque international : Contraintes linguistiques, linguistique contrainte. A propos de la complémentation nominale. Université Paris-Descartes 2-3 juin 2016.

Blasco, M. & Cappeau, P. (2012). Identifier et caractériser un genre : l'exemple des interviews politiques, Langages 187, 27-40.

Caron, J. et coll. (1983). La pensée naturelle, structures, procédures et logique du sujet. Groupe de Recherche Ontogénèse des Processus Psychologiques. Université de Rouen : Publication de l'Université de Rouen, 86, Paris : PUF.

Caron, J. (1979). Compréhension d'un connecteur polysémique : la conjonction "si", Bulletin de psychologie, N° spécial "compréhension du langage" 32, 791-801.

Chaurand, J. & Mazière, F. (1990). La définition. Centre d'études du lexique. Paris : Larousse, 97-110.

Clark, E. V. (2003). First Language Acquisition. Cambridge : Cambridge University Press. François, F. (1981). « Exemples de maniement "complexe" du langage : définir-résumer », in APREF, J'cause français, non ? Paris : La découverte, Maspero.

François, F. (1985). Qu'est-ce qu'un ange ? Ou définition et paraphrase chez l'enfant, in C. Fuchs (éds.). Aspects de l'ambiguité et de la paraphrase dans les langues naturelles. Bern : P. Lang.

Grize, J.B. (1990). Logique et langage. Paris : Ophrys.

Lebas-Fraczak, L. (2016). Les opérations intellectuelles des élèves et la perception de l'enseignante dans trois discussions philosophiques en classe de CP, Recherche En Education, 25, 133-146.

Leleux, C. (éd.). (2005). La philosophie pour enfants. Le modèle de Matthew Lipman en discussion [Philosophy for children. A discussion of the model of Matthew Lipman]. Bruxelles : De Boeck / Larcier.

Lipman, M./Sharp, A.-M. & Oscanyan, F. S. (1980). Philosophy in the classroom. Philadelphia, PA : Temple University Press.

Martin, R. (1990). La définition naturelle, in : J. Chaurand & F. Mazière. La définition. Centre d'études du lexique. Paris : Larousse, 86-95.

Nonnon E. (1996). Activités argumentatives et élaboration de connaissances nouvelles : le dialogue comme espace d'exploration, Langue Française 112, 67-87.

Nonnon E. (1999). L'enseignement de l'oral et les interactions verbales en classe : champs de référence et problématiques, Revue française de pédagogie 129, 87-131.

Nonnon, E. (1990). Mouvements discursifs et modes de réflexion en commun dans des discussions d'adolescents en échec scolaire. Thèse de 3ème cycle. Paris V.

Nonnon, E. (1991). Mettre en tableau, mettre au tableau ou comment structurer les discussions d'enfants ? Logique naturelle à l'oral et formalisations écrites, L'oral dans l'écrit. E.L.A. 81, 95-117.

Rastier, F. 1991. Sémantique et recherches cognitives. Paris : PUF.

Revault d'Allonnes, M. & Foessel, M. (dir.) (2012). Chouette ! Philo. Abécédaire d'Artiste à Zombie. Paris : Gallimard Jeunesse.

Rey-Debove, J. (1988). Prototypes et définitions, *DRLAV* (Centre de linguistique et de recherche de l'Université de Paris 8), 41, 143-167.

Rey-Debove, J. (1989). Dictionnaires d'apprentissage : que dire aux enfants ?, *Le français dans le monde* (numéro spécial). *Lexiques*, 18-23.

Rey-Debove, J. (1993). Le contournement du métalangage dans les dictionnaires pour enfants : translation, monstration, neutralisation, *Repères* 8, 79-92.

Rey-Debove, J. (1999). *La linguistique du signe.* Paris : Colin.

Riegel, M. (1987). Définition directe et indirecte dans le langage ordinaire : les énoncés définitoires copulatifs, *Langue française* 73-1, 29-53.

Riegel, M. (1990). *La définition*, acte du langage ordinaire. De la forme aux interprétations, in J. Chaurand & F. Mazière (éds.). La définition. Centre d'études du lexique. Paris : Larousse, 97-110.

Rispail, M. (2007a). Le modèle américain : Matthew Lipman, in : M. Rispail (dir.). *Apprendre à parler, apprendre à penser. Les ateliers de philosophie.* Paris : Sceren, Nice : CRDP, 23-24.

Rispail, M. (2007b). Les modèles français : Jacques Lévine et Michel Tozzi, in M. Rispail (dir.). *Apprendre à parler, apprendre à penser. Les ateliers de philosophie.* Paris : Sceren, Nice : CRDP, 25-27.

Rispail, M. (2007c). Rôle de l'enseignant dans l'atelier de philosophie, in M. Rispail (dir.). *Apprendre à parler, apprendre à penser. Les ateliers de philosophie.* Paris : Sceren, Nice : CRDP, 77-98.

Roiné, P. (2015). Étude des éléments dialogiques prÉsents dans les processus de conceptualisation lors des Discussions à Visée Philosophique en cycle 3 de l'École élémentaire. Thèse de Doctorat en Sciences du Langage. Paris-Créteil : Université de Cergy Pontoise.

Rossi, M. (2007). *Déc(rire) le monde : formes de comique involontaire dans les définitions spontanées des enfants,* Bouquets pour Hélène, 6, http://publifarum.farum.it/ezine_printarticle.php?id=30

Sharp, A. M. (1990). La communauté de recherche : une éducation pour la démocratie, in A. Caron (éd.). *Philosophie et pensée chez l'enfant.* Montréal : Agence d'Arc, 85-103.

GRECO L. & TRAVERSO V. (éds.) (2016), *Langages : Définir les mots dans l'intercation : un essai de sémantique interactionnelle,* 204, Paris : Larousse.

Topping, K. J., & Trickey, S. (2007). Collaborative philosophical enquiry for school children : Cognitive effects at 10–12 years, *British Journal of Educational Psychology,* 77, 271–288.

Trognon A., (1999). Eléments d'analyse interlocutoire, dans Gilly M., Roux J.P., Trognon A., *Apprendre dans l'interaction* (pp.69-94). Nancy Aix-en-Provence : Publication de l'Université de Provence. Nancy..

Trognon A. (1995). La fonction des actes de langage dans l'interaction : l'exemple de l'intercompréhension en conversation, *Lidil, Revue de linguistique et de didactique des langues, L'interaction en question, 12*, 67-85.

Wierzbicka A. (1996). *Semantics : Primes and Universals.* Oxford : University Press.

Étude quantitative des propriétés dialogiques des adverbes épistémiques

Corinne Rossari et Margot Salsmann
Université de Neuchâtel
prenom.nom@unine.ch

1. Introduction

Dans cette communication, nous allons utiliser une approche de linguistique outillée de corpus pour voir si l'on parvient à dégager des régularités qui rendent compte du fait que certains adverbes épistémiques (à savoir ceux qui ont la propriété de jeter un doute, même minime, sur l'état de choses) ont développé des usages non épistémiques de type concessif, comme dans les exemples suivants :

- (1) 14 novembre 2015 : Je suis peut-être suisse, mais je soutiens totalement la France. (https://www.youtube.com/watch?v=rjF-La45)
- (2) Je suis certes une femme, mais je m'impose. Dans notre travail, il y a toujours eu des actes de violence, c'est le côté pénible de la profession. (http://www.5plus.mu/node/17279)

La propriété de nationalité ou de genre, appliquée au locuteur, n'est pas passible d'une évaluation épistémique (voir Rossari et al. 2016). Or *peut-être* et *certes* peuvent figurer dans ce type de contexte.

- (3) ? ?Je suis certainement/probablement... une femme, mais je m'impose. Dans notre travail, il y a toujours eu des actes de violence, c'est le côté pénible de la profession.

Ces deux adverbes émettent des jugements épistémiques bien distincts. Si *certes* engage le locuteur vers une appréciation qui évalue positivement la réalisation d'un état de choses, il n'en va pas de même pour *peut-être*, qui met en doute ce dernier. Des contextes dialogiques comme les suivants mettent en relief les évaluations épistémiques presque antagonistes des deux adverbes :

- (4) Est-ce que Pierre est là ?
  Certes, le voici ! / ? ?Peut-être, le voici !

Certes à lui seul peut être utilisé pour confirmer un état de choses, fonction que ne peut avoir *peut-être* qui, de ce fait, n'est pas naturel avec une forme déictique (*le voici*) qui atteste la présence de Pierre.

2. Utilisation de la linguistique outillée de corpus

Notre corpus de départ est constitué de textes issus de la presse écrite contemporaine : Le Monde 2008 (20 410 766 mots), Le Figaro 2008 (10 795 373 mots) et Sud-Ouest 2002 (29 763 988 mots). Ces trois quotidiens représentent deux journaux nationaux d'orientation politique différente et un journal régional. Comme nous le verrons par la suite, ces corpus font état de comportements très homogènes concernant nos adverbes. Nous avons alors ajouté à ceux-ci des corpus relevant de genre et d'époque différents : l'Encyclopaedia Universalis 2005 (49 859 864 mots), Wikipédia 2015 (50 396 345 mots) et l'Encyclopédie de Diderot et d'Alembert, édition 1751-1772 (23 940 181 mots). Ces corpus font apparaître une disparité sensible quant au comportement de ces adverbes, ce qui nous conduira, en conclusion, à nous interroger sur l'incidence du genre et de l'époque sur le comportement de ces adverbes. Ces corpus sont fournis par la plateforme BTLC [1] (cf. Diwersy 2014) – que nous exploitons

1. Les corpus utilisés sont issus du projet PRESTO (http://presto.ens-lyon.fr). Le corpus de référence du projet franco-allemand PRESTO pour la période XVIe s. – XXe s. a été constitué grâce aux textes issus des bases textuelles suivantes : FRANTEXT (http://www.frantext.fr, V. Montémont, G. Souvay), BVH (Bibliothèques Virtuelles Humanistes, http://www.bvh.univ-tours.fr, L. Bertrand, M.-L. Demonet), ARTFL (American and French Research on the Treasury of the French Language, http://artfl-project.uchicago.edu, R. Morrissey, M. Olsen) et CEPM (Corpus électronique de la première modernité, http://www.cpem.paris-sorbonne.fr). Les ressources et les outils élaborés dans PRESTO ont bénéficié des apports des logiciels LGerM (lemmatisation de la variation graphique des états anciens du français et lexiques morphologiques, http://www.atilf.fr/LGeRM, G. Souvay) et Analog (M.-H. Lay), ainsi que du lexique Morphalou (http://www.cnrtl.fr/lexiques/morphalou).

pour l'extraction de nos données.

Le paradigme d'adverbes épistémiques que nous prenons en considération est fondé sur la classification de Molinier et Lévrier (2000 : 92). À partir de la liste des adverbes modaux qu'ils ont constituée, nous avons retenu ceux qui instaurent un doute, aussi minime soit-il, sur l'état de choses, à savoir : *peut-être, sans doute, certes, probablement, certainement, sûrement, vraisemblablement, assurément.*

Notre but est d'identifier des régularités dans le comportement des adverbes épistémiques à potentiel concessif en portant notre attention (i) sur la fréquence de ces adverbes dans chacun de ces corpus ; (ii) sur leurs cooccurrents spécifiques ; (iii) sur les places syntaxiques qu'ils sont susceptibles d'occuper.

(i) Les fréquences relatives de ces adverbes nous permettent de voir si le fonctionnement concessif de certains d'entre eux est lié au fait qu'ils sont plus ou moins fréquents dans les corpus.

(ii) En ce qui concerne les cooccurrents spécifiques, nous testons ceux qui sont susceptibles de donner à ces adverbes une valeur concessive, à savoir, d'un côté, leur aptitude à être associé au « ?» juste avant leur occurrence (empan 1), pour s'assurer que l'adverbe est utilisé en tant que réaction à une question préalable, et, de l'autre, leur aptitude à être associé à *mais* dans un empan de max. 10 items après l'adverbe, pour réduire la possibilité que l'adverbe et l'emploi de *mais* ne soient pas en relation. Ces deux tests reposent sur l'idée que les adverbes à emploi concessif sont fondamentalement dialogiques (cf. Rossari 2016) et sur l'idée que leur emploi non épistémique (dit concessif) est lié à leur fréquence élevée dans l'entourage de *mais*.

(iii) Le calcul de la fréquence concernant la place syntaxique de l'adverbe dans l'énoncé nous permettra d'évaluer leur potentiel dialogique d'une autre façon. En effet, la place initiale est nettement préférée pour des formes prototypiquement réactives comme *oui* et *non* (nous renvoyons aux travaux de Borillo 1976 et Molinier et Lévrier 2000 pour la propriété qu'ont les adverbes épistémiques de constituer une réponse à une question totale).

Enfin, les résultats obtenus seront mis en perspective avec le genre et l'époque propres à chaque corpus.

3. Résultats sur la fréquence relative et les cooccurrents spécifiques

(i) Les fréquences relatives (cf. tableaux 1 et 2) de chacun de ces adverbes dans les six corpus indiquent que la répartition des adverbes paraît être sensible aux genres et aux époques. En effet, dans les corpus de presse – Le Monde (LM), Le Figaro (LF) et Sud-Ouest (SO) –, la distribution est très semblable. L'ordre dans lequel apparaissent les adverbes est le même et leur fréquence relative reste sur une échelle similaire, par ex. *peut-être* (LM : 142 (par mio de mots) / LF : 171 / SO : 166). Les deux corpus encyclopédiques contemporains donnent des résultats plus contrastés. Dans l'Encyclopaedia Universalis (EU), on retrouve en tête le même trio que dans LM, LF et SO, mais pas dans le même ordre (*sans doute, peut-être, certes* pour EU et *peut-être, sans doute, certes* pour LM, LF et SO). Dans Wikipédia (WI), l'ordre et les fréquences sont très différents : les fréquences sont nettement plus basses pour tous les adverbes et le trio de tête est constitué par *probablement, peut-être* et *sans doute*. Certes arrive en 6ème position avec une fréquence relative bien inférieure à celle qu'il a dans les autres corpus : 10 (WI) vs. 84 (EU), 93 (SO), 100 (LF) et 147 (LM). Le corpus du XVIIIe, l'Encyclopédie de Diderot et d'Alembert (DDA), donne un autre classement encore : *peut-être, sans doute, certainement* constituent le trio de tête, et *certes* arrive en fin de classement avec une fréquence particulièrement faible (1,46).

(ii) Pour l'étude des cooccurrents spécifiques, nous avons ensuite appliqué la mesure d'association « log-likelihood » (LL) (cf. Evert 2008) pour déterminer le rang auquel « ? » et *mais* interviennent en tant que cooccurrent spécifique de chaque adverbe (cf. lexicogrammes 1 et 2) et pour comparer les adverbes entre eux en fonction de leur force d'attraction avec le « ? » et *mais* (cf. tableaux 1 et 2). Cette mesure donne la probabilité selon laquelle l'association d'un item avec un cooccurrent n'est pas due au hasard. Nous avons retenu le seuil qui permet d'avoir 99% de probabilité que l'association soit non aléatoire, à savoir 10,83. Nous commentons ici uniquement les résultats du corpus journalistique.

Nous présentons en annexe les lexicogrammes des quatres adverbes les plus fréquents : *certes, peut-être, sans doute, probablement* (cf. lexicogrammes 1 et 2). En ce qui concerne le « ? », les lexicogrammes ne permettent pas de faire ressortir les deux adverbes à emploi concessif par rapport aux autres : le

« ? » est spécifique pour tous les adverbes (à des degrés divers) sauf pour *vraisembablement*. Le rang auquel intervient *mais* apporte en revanche un éclairage plus pertinent. *Peut-être* et *certes* ont ce connecteur comme premier cooccurrent spécifique dans tous les corpus de presse ; s'ajoute à ce duo, *sans doute* et *certainement*, pour lesquels *mais* arrive respectivement dans les cinq et sept premiers cooccurrents. Il n'apparaît ni pour *vraisemblablement*, ni pour *probablement* dans aucun des corpus.

Si on se réfère au tableau comparatif entre adverbes (cf. tableaux 3 et 4), on relève des contrastes entre *peut-être, certes* et *sans doute* et les autres adverbes. En effet, ces trois adverbes cumulent les valeurs les plus élevées pour les deux cooccurrents : « ? » et *mais*. On remarque aussi que les adverbes dont la valeur avec le « ?» est la plus faible n'ont pas *mais* comme cooccurrent spécifique.

Ces premiers résultats font ressortir c*ertes, peut-être* et, dans une moindre mesure, *sans doute*. Ce dernier est également compatible avec une lecture concessive, même si son emploi paraît moins naturel dans les exemples (1) et (2). On trouve toutefois des occurrences de ce type :

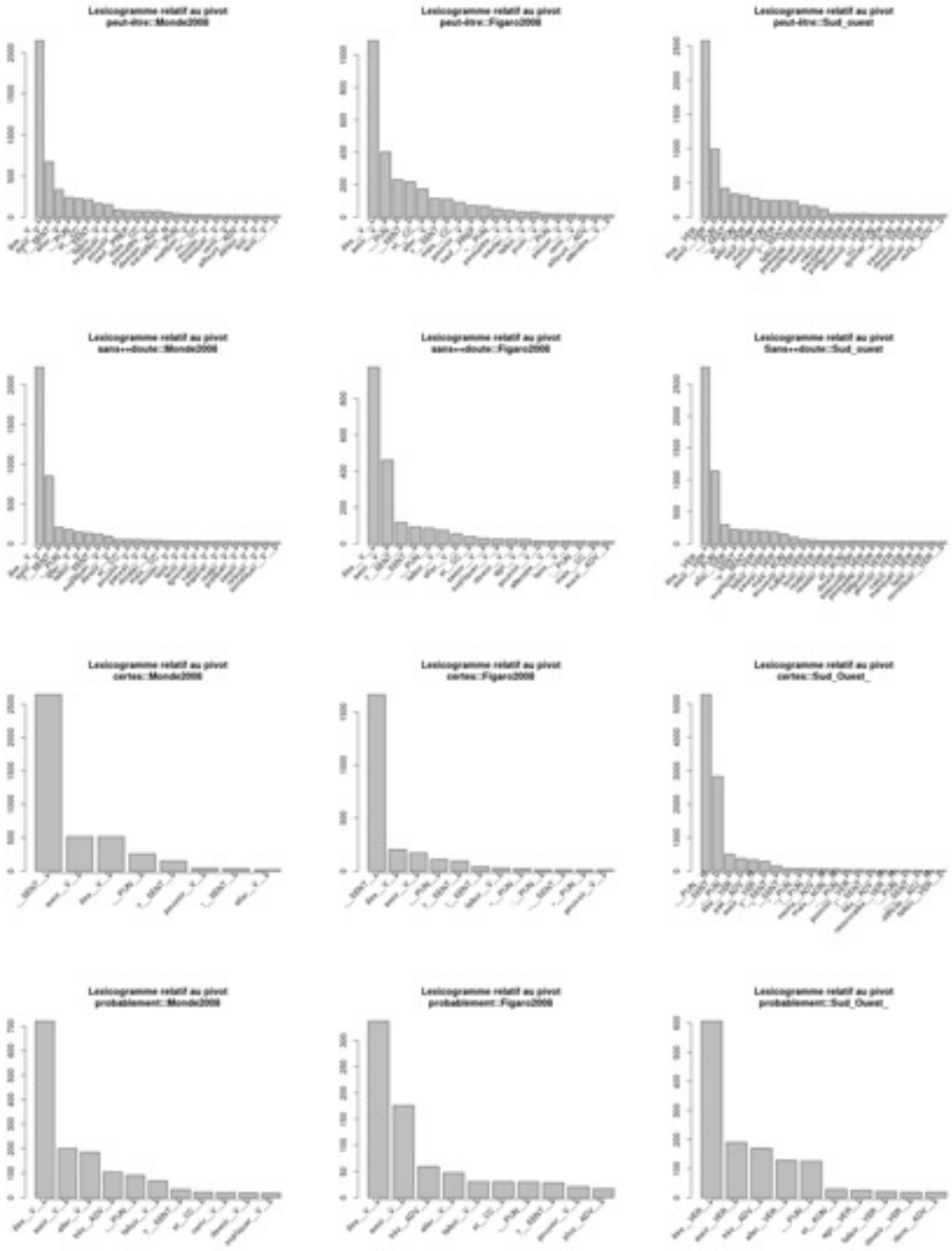
- (5) Je suis sans doute une femme mûre, mais je m'entretiens et je suis fière de mon corps. (www.deadly-sexe.com/recits_le-sauna.html)

Les requêtes sur la position des adverbes permettront de croiser ces premiers résultats avec un tout autre paramètre. On verra ainsi s'il y a une convergence entre l'attraction avec « ? » comme cooccurrent gauche, avec « mais» comme cooccurrent droit et la position initiale (prise comme un indice de fonctionnement dialogique).

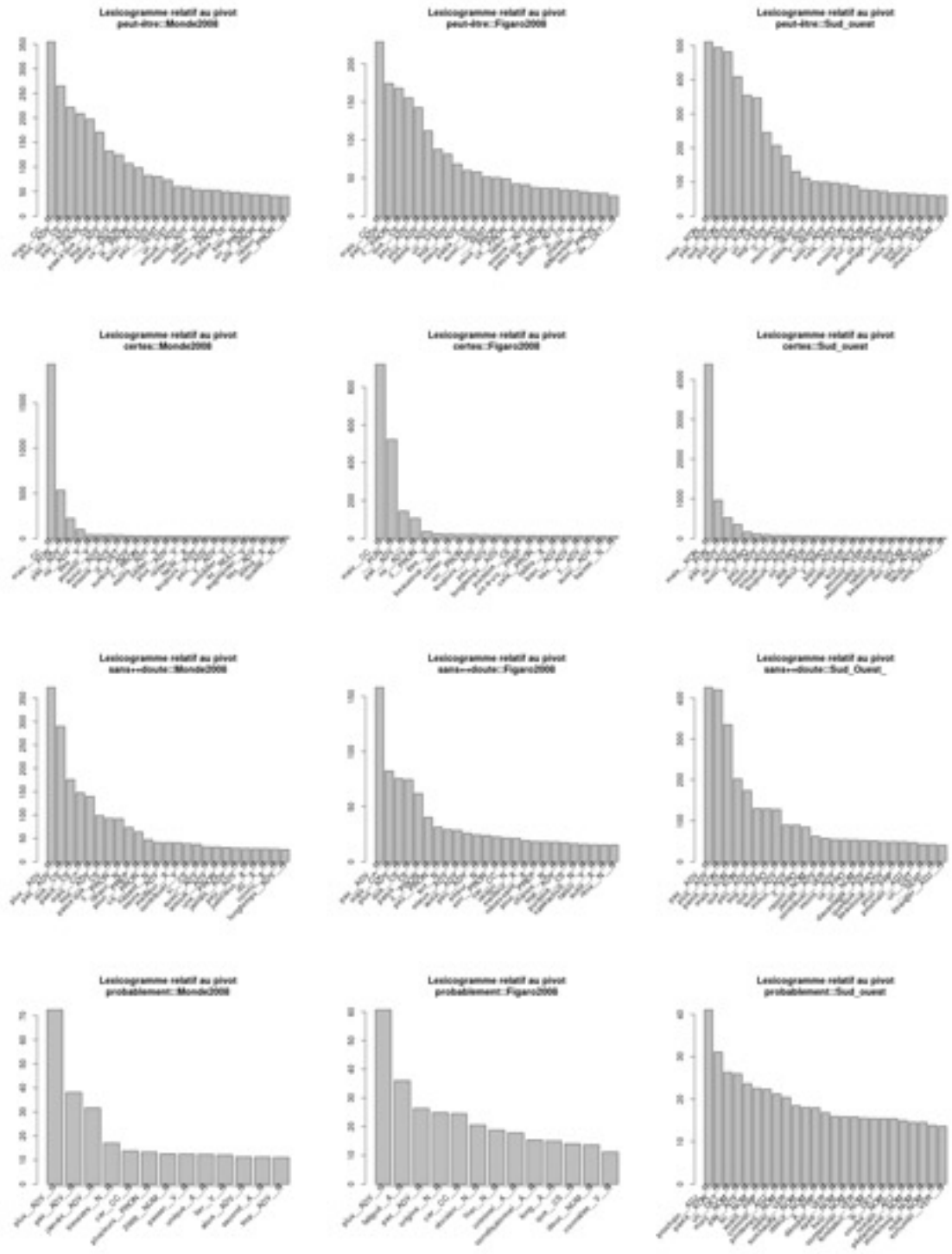
Les résultats que nous avons commentés devront être mis en perspective avec les corpus encyclopédiques du XXIe et du XVIIIe, qui présentent une image très différente des attractions entre ces adverbes et les deux cooccurrents pris en compte, nous conduisant à nous interroger sur l'incidence du genre sur les valeurs observées.

4. Lexicogrammes et tableaux

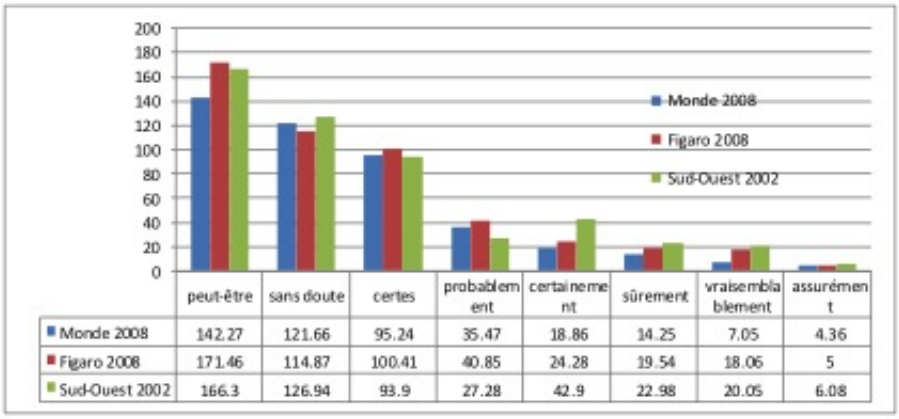
- Lexicogrammes 1 : cooccurrences spécifiques à gauche (empan 1)



- Lexicogrammes 2 : cooccurrences spécifiques à droite (empan 10)

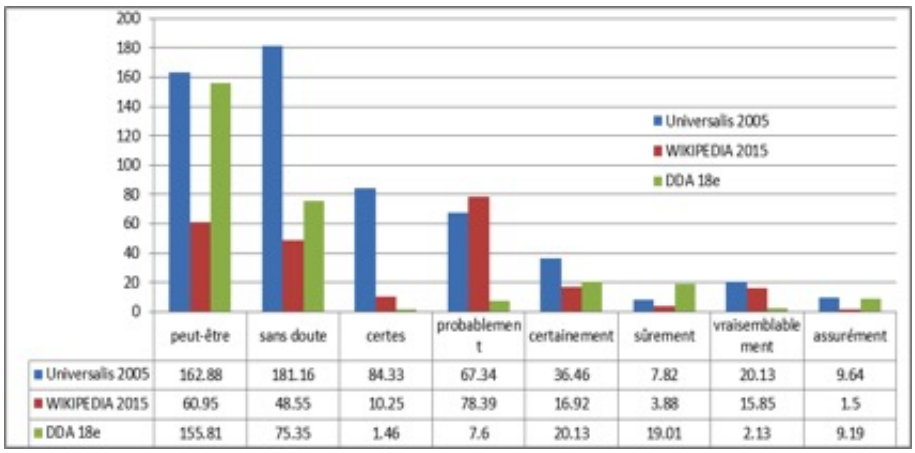


- Tableau 1 : fréquences relatives des adverbes (par mios de mots) – Corpus journalistique



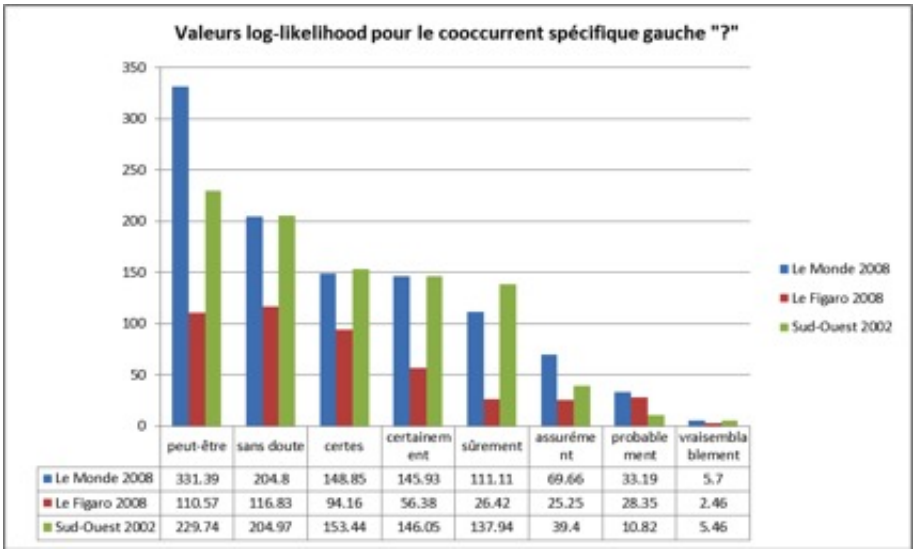
	peut-être	sans doute	certes	probablement	certainement	sûrement	vraisemblablement	assurément
Monde 2008	142.27	121.66	95.24	35.47	18.86	14.25	7.05	4.36
Figaro 2008	171.46	114.87	100.41	40.85	24.28	19.54	18.06	5
Sud-Ouest 2002	166.3	126.94	93.9	27.28	42.9	22.98	20.05	6.08

- Tableau 2 : fréquences relatives des adverbes (par mios de mots) – Corpus encyclopédique



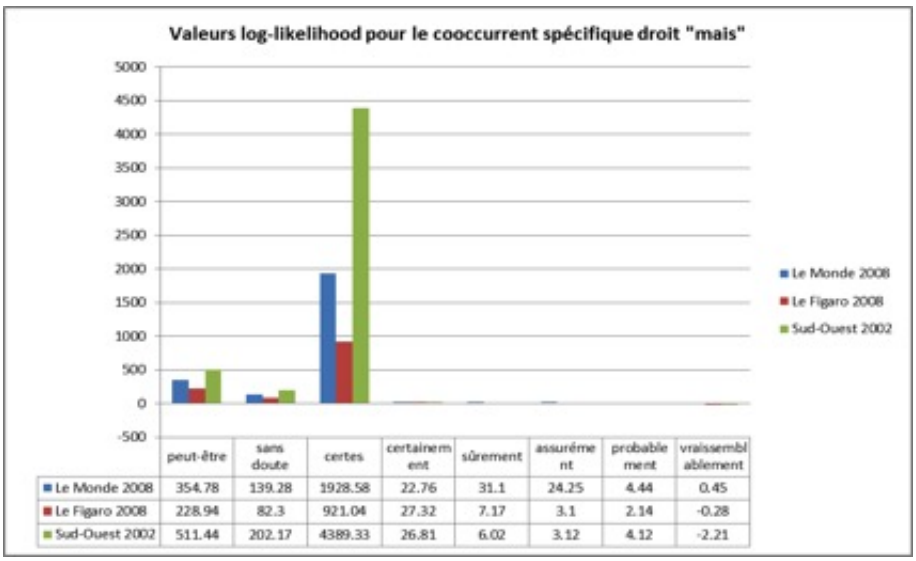
	peut-être	sans doute	certes	probablement	certainement	sûrement	vraisemblablement	assurément
Universalis 2005	162.88	181.16	84.33	67.34	36.46	7.82	20.13	9.64
WIKIPEDIA 2015	60.95	48.55	10.25	78.39	16.92	3.88	15.85	1.5
DDA 18e	155.81	75.35	1.46	7.6	20.13	19.01	2.13	9.19

- Tableau 3 : valeurs log-likelihood (LL) pour le cooccurrent spécifique gauche « ? »



	peut-être	sans doute	certes	certainement	sûrement	assurément	probablement	vraisemblablement
Le Monde 2008	331.39	204.8	148.85	145.93	111.11	69.66	33.19	5.7
Le Figaro 2008	110.57	116.83	94.16	56.38	26.42	25.25	28.35	2.46
Sud-Ouest 2002	229.74	204.97	153.44	146.05	137.94	39.4	10.82	5.46

- Tableau 4 : valeurs log-likelihood (LL) pour le cooccurrent spécifique droit mais



	peut-être	sans doute	certes	certainement	sûrement	assurément	probablement	vraisemblablement
Le Monde 2008	354.78	139.28	1928.58	22.76	31.1	24.25	4.44	0.45
Le Figaro 2008	228.94	82.3	921.04	27.32	7.17	3.1	2.14	-0.28
Sud-Ouest 2002	511.44	202.17	4389.33	26.81	6.02	3.12	4.12	-2.21

Références bibliographiques

Blumenthal, P. (2011). « Odeur – évolution des profils combinatoires», *Langages, n°181*, 53-71.

Borillo, A. (1976). « Les adverbes et la modalisation de l'assertion», *Langue française, n°30*,74-89.

Diwersy, S. (2014). « La plateforme Primestat. BTLC et l'exploitation lexico-statistique de corpus diachroniques », communication présentée à la Journée d'études organisée à l'Institut des Sciences du Langage et de la Communication, Université de Neuchâtel.

Evert, S. (2008). « Corpora and collocations », in Lüdeling, A., Kytö, M. (éds), *Corpus Linguistics. An International Handbook*, Berlin : Mouton de Gruyter, 1212-1248.

Molinier, C. et Lévrier, F. (2000). *Grammaire des adverbes. Description des formes en -ment*, Genève-Paris : Droz.

Rossari, C. (2016), « La concession sans opposition à la lumière de la théorie argumentative de la polyphonie », *Verbum, XXXVIII*, 151-168.

Rossari, C., Hütsch, A., Ricci, C., Salsmann, M. et Wandel, D. (2016) « Le pouvoir attracteur de *mais* sur le paradigme des adverbes épistémiques : du quantitatif au qualitatif », Actes des *13èmes Journées internationales d'Analyse statistique des Données Textuelles* (en ligne : https://jadt2016.sciencesconf.org/82999).

Session 5.B.
Français-Allemand

Apprivoiser les virgules en allemand – Une approche sur corpus

Eva Schaeffer-Lacroix
Université Paris-Sorbonne / Espé de Paris
Laboratoire CeLiSo (Centre de Linguistique en Sorbonne)
evalacroix@free.fr

1. Introduction

L'usage de corpus numériques pour divers aspects de l'apprentissage d'une langue étrangère commence à être une pratique relativement courante. Ädel (2010), Landure & Boulton (2010) et O'Sullivan (2010) s'intéressent plus particulièrement au rôle des corpus pour l'apprentissage de la production écrite. Boch et Frier (2015) et Chachkine, Demaizière & Schaeffer-Lacroix (2013) font un lien explicite entre exploration de corpus, révision de texte et réflexion sur la langue. Après avoir exploré l'enseignement-apprentissage du sens de particules verbales allemandes à l'aide d'un corpus contenant un script de film (Schaeffer-Lacroix, 2016), je m'intéresse actuellement à la méthode d'exploration de corpus pour soutenir la conceptualisation de l'usage de la virgule dans des écrits académiques, plus précisément dans des comptes-rendus de lecture d'articles allemands.

Pourquoi les apprenant·e·s d'allemand langue étrangère devraient-ils chercher à "apprivoiser" [1] la virgule allemande, objet d'apparence insignifiante et négligeable, dans les textes qu'ils produisent ? On pourrait estimer qu'il y a des problèmes de formulation plus saillants. Toutefois, la virgule contribue à la structuration du texte : elle renseigne sur le lien entre ses différents éléments ; elle signalise des frontières à l'intérieur des phrases et elle marque des relations anaphoriques. Selon Favriaud (2011 : 3),

> la ponctuation pose toujours (...) la question de la communication écrite, de la distinction et de l'organisation des parties du discours, de la lisibilité de celui-ci et des effets produits sur les récepteurs. La ponctuation a une vertu réflexive et réfléchissante en regard de l'écrit.

On peut donc conclure qu'en situation d'apprentissage de la production écrite, l'insertion (ou non) des virgules est un domaine qui renseigne sur certaines compétences du scripteur, en l'occurrence son degré de compréhension du fonctionnement de la langue. Cette constatation permet de considérer la gestion de la virgule par les scripteurs plus ou moins novices en allemand comme un objet d'étude valable.

Kirchhoff et Primus (2016 : 78ff) proposent une analyse multilingue du phénomène : ils comparent l'emploi de virgules dans cinq situations différentes en allemand, anglais, espagnol et russe. Dans deux des situations, des différences notables entre les langues peuvent être observées. La première différence concerne ce que Boettcher (2016 : 336) appelle le "marquage de territoire" [2] entre principale et subordonnée par une virgule en allemand et en russe (Kirchhoff & Primus, 2016 : 87) : "Ich glaube, dass niemand zur Party gekommen ist." respectivement "Я полагаю, что никто не прибыл на вечеринку." [Je crois que personne n'est venu à la fête."]. Ni en anglais ni en espagnol (pas plus qu'en français, d'ailleurs), la virgule n'est utilisée dans de tels contextes : "I think that no one came to the party."/"Creo que no fue nadie a la fiesta.". La deuxième différence concerne la gestion des groupes infinitifs qui est assez complexe en allemand, contrairement aux autres langues analysées par Kirchhoff et Primus (2016 : 88). Ces deux particularités de la langue allemande fournissent un cadre pertinent pour l'analyse d'erreurs de virgule dans des corpus d'apprenants. De plus, ces deux caractéristiques sont susceptibles d'offrir à des apprenants scripteurs (Leblay, 2014) des pistes pour la formulation d'hypothèses lors de l'exploration de la virgule dans des corpus contenant des textes de scripteurs experts, aussi appelés "scripteurs expérimentés" (Leblay, 2009), traitant un sujet comparable. Cette exploration peut ensuite être mise à profit pour la révision de texte.

1. Définition d'"apprivoiser" dans le *Trésor de la Langue Française Informatisé* : "Apprivoiser qqn à, avec qqc. Le familiariser progressivement avec quelque chose, lui faire abandonner son hostilité ou ses réticences... (...) Dominer, maîtriser progressivement (...)."

2. Dans ses termes : "*die Territorien der beteiligten Prädikate (...) markieren*" (Boettcher, 2016 : 337).

2. Contexte

L'équipe master MEEF (métiers de l'enseignement, de l'éducation et de la formation), parcours allemand, de l'Espé de Paris m'a confié le suivi de l'activité de production écrite en allemand langue étrangère d'étudiant·e·s participant à un module dédié aux courants majeurs en didactique. Voici la démarche adoptée : avant chaque rencontre en présentiel, les étudiant·e·s doivent lire un article en langue allemande portant sur le thème de la séance à venir. Au moins un des membres du groupe en fait un compte-rendu, et il le dépose sur l'espace numérique de travail du module afin que son texte puisse être lu et commenté par les pairs. Je procède ensuite à une discrète annotation – par exemple, surlignage de parties erronées selon un code couleur – du compte-rendu et des commentaires afin de signaler aux apprenants scripteurs les endroits méritant une révision.

Lors de la phase d'annotation des textes des non-germanophones de la promotion actuelle, j'ai observé que, dans pratiquement toutes les productions, les virgules ont tendance à être absentes ou à apparaître à des endroits non-pertinents. Cette constatation trouve un écho dans les "Indications aux candidat(e)s quant à l'orthographe allemande", concernant plus particulièrement la virgule, que les auteurs du rapport de jury du CAPES d'allemand de la session 2016 ont fournies aux candidat·e·s afin de les aider à améliorer leur production écrite lors de la phase d'admissibilité du concours (Goullier, 2016 : 72) [3].

3. Méthodes

Lors de la dernière séance du module de didactique, j'ai proposé aux deux groupes de la promotion actuelle (n=9 resp. n=5) un atelier corpus dédié à la prise en main du système de gestion de corpus *Sketch Engine* (Kilgarriff, Rychly & Pomikalek, nd) afin qu'ils puissent s'en servir dans une visée linguistique et didactique (cf. *learner as a researcher* dans Bernardini, 2004 :15). La formation linguistique a été explicitement articulée à la formation technique : lors de l'introduction à *Sketch Engine*, la virgule a fait l'objet d'illustration de types de requête. Les étudiant·e·s avaient l'occasion de repérer des occurrences de virgules dans un corpus d'experts (Boch & Frier, 2015 : 91) qui contient des comptes-rendus rédigés par les germanophones d'une session antérieure. Après une phase d'exploration de ce corpus en binômes, nous avons procédé à un échange oral en commun afin de classer les occurrences, observées par les binômes, dans les catégories proposées par Boettcher (2016 : 330-338). Ces catégories permettent de réduire à quatre les 132 règles concernant la virgule citées dans le *Duden online* (nd), outil de référence pour les normes de la langue allemande. Les scripteurs avaient ensuite l'occasion d'explorer le corpus d'experts dans le but de réviser leurs textes du point de vue des virgules, mais aussi pour améliorer les autres passages annotés. L'intégralité des deux ateliers corpus (deux fois 6 heures) a été enregistrée, et les écrans ont été filmés lors des explorations de corpus. Le projet s'est terminé par un entretien filmé proposé à chacun des deux groupes.

4. Analyse des données

Quatre types de révision ont pu être identifiés dans les données : révision pertinente, révision non effectuée, révision non pertinente et formulation d'une question de travail par les scripteurs.

Révision pertinente et révision non effectuée

La Figure 1 illustre la prise en compte d'une suggestion d'ajout d'une virgule par Laura [4], mais on peut également constater que le deuxième endroit qui mériterait l'insertion d'une virgule (signalé par "xxx") n'a pas été modifié.

3. Voir aussi le rapport du CAPLP Externe et CAFEP-PLP, section Langues vivantes-Lettres, option Allemand, session 2016. On y déplore "l'emploi erratique voire l'omission de la virgule dans un très grand nombre de copies" et on y précise qu'"en allemand, les virgules sont essentielles pour lire et comprendre le texte" (Erin, 2016 : 10).

4. Les prénoms des participant·e·s ont été anonymisés.

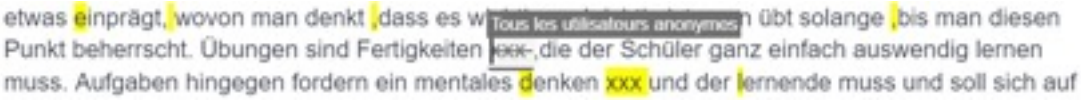


FIGURE 1 – **Ajout d'une virgule par Laura.**

Révision non pertinente

Dans l'exemple ci-dessous, Agnès place un complément prépositionnel (mis en gras par moi) entre virgules et le traite ainsi comme une incise.

> Die Handlungs- und Aufgabenorientierung ist, **in der Fremdsprachendidaktik**, ein sehr wichtiger Begriff [5]. [L'orientation actionnelle et par tâches est, en didactique des langues étrangères, un terme très important .]

Il y a des situations où il est légitime d'entourer un complément prépositionnel de virgules, mais en l'occurrence, il s'agit d'un élément constitutif du sens de l'énoncé ce qui fait que les virgules ne sont pas requises en allemand. Un test de déplacement du complément en est l'illustration [6].

> **In der Fremdsprachendidaktik** ist die Handlungs- und Aufgabenorientierung ein sehr wichtiger Begriff.

Formulation d'une question de travail

Sarah s'est interrogée au sujet de la nécessité de placer une virgule avant la conjonction de subordination *wie* [comment] dans l'extrait ci-dessous. Elle a visiblement du mal à la distinguer de la conjonction de coordination homographe qui n'est pas précédée d'une virgule.

Comme on peut le voir en Figure 2, Sarah a formulé une question permettant de prolonger le débat amorcé avec les pairs avant la phase de révision.

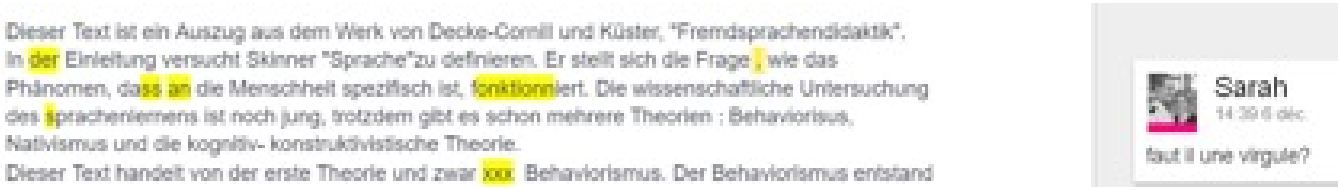


FIGURE 2 – **Formulation d'une question de travail par Sarah**

5. Conclusion

La recherche qui vient d'être esquissée a abouti à des résultats concluants dans plusieurs domaines : les enregistrements des écrans permettent d'observer le degré de convergence entre questions de travail linguistiques et requêtes sur corpus. Les traces des échanges oraux en groupe entier contiennent des indices d'une meilleure compréhension du rôle de la virgule en allemand. L'analyse de la façon dont les étudiant·e·s ont révisé leurs textes permet d'identifier plusieurs types d'actions, dont la continuation de l'échange autour du rôle de la virgule en allemand. De plus, les scripts des entretiens permettent de constater la large adhésion des participant·e·s au choix de la méthode de révision à l'aide de *Sketch Engine* (Kilgarriff *et al.*, nd). On peut donc conclure que l'architecture générale du scénario et le choix des outils conceptuels et techniques proposés aux deux groupes sont pertinents et méritent d'être reproposés lors d'actions de formation à venir.

6. Du point de vue de la structure informationnelle, j'aurais d'ailleurs tendance à préférer l'ordre syntaxique du test de déplacement : complément prépositionnel – verbe – sujet – attribut du sujet.

Références bibliographiques

Ädel, A. (2010). Using corpora to teach academic writing : challenges for the direct approach. In Mari Carmen Campoy-Cubillo, Begoña Bellés-Fortuño & Maria Lluisa Gea-Valor (dir). *Corpus-based approaches to English language teaching*. Londres, New York : Continuum, 39-55.

Bernardini, S. (2004). Corpora in the classroom. An overview and some reflections on future developements. In Sinclair, J. (dir.). *How to Use Corpora in Language Teaching*. Amsterdam, Philadelphia : John Benjamins Publishing Company, 15-36.

Boch, Françoise & Frier, Catherine (2015). Travailler le texte : ponctuation, anaphores et collocations. In Françoise Boch & Catherine Frier (dir). *Écrire dans l'enseignement supérieur : Des apports de la recherche aux outils pédagogiques*. Grenoble : UGA Éditions (Didaskein), 53-109.

Boettcher, Wolfgang (2016). Komma & Co unter dem Kooperationsprinzip : Interpunktionslernen im Kompetenzbereich 'Schreiben und Sprachreflexion' [Virgule & co sous le principe de la coopération : apprentissage de la ponctuation dans le domaine de compétence 'Écrire et réflexion sur la langue']. In Ralph Olsen, Christiane Hochstadt & Simona Colombo-Scheffold (dir.). *Ohne Punkt und Komma ... Beiträge zu Theorie, Empirie und Didaktik der Interpunktion*. Berlin : RabenStück Verlag, 326-361.

Chachkine, Elsa, Demaizière, Françoise & Schaeffer-Lacroix, Eva (2013). "Pour un apprenant réfléchissant". Linguistik online 60, 3/2013, 23-42. (2013). Chachkine, E., Demaizière, F. & Schaeffer-Lacroix, E. "Pour un apprenant réfléchissant". *Linguistik online 60, 3/2013*, 23-42.

Duden online (nd). Dictionnaire électronique. Entrée "*Komma*" [Virgule]. Berlin : Bibliographisches Institut GmbH. http://www.duden.de/sprachwissen/rechtschreibregeln/komma

Erin, Jonas (2016). *Rapport de jury CAPLP Externe et CAFEP-PLP, section Langues vivantes-Lettres, option Allemand, session 2016*. Ministère de l'Éducation Nationale, de l'Enseignement supérieur et de la Recherche. http://media.devenirenseignant.gouv.fr/file/externe/21/1/rj-2016-CAPLP-externe-Allemand-Lettres_632211.pdf

Favriaud, Michel (2011). Approches nouvelles de la ponctuation, diachroniques et synchroniques. *Langue française*, 4/2011, n°172, 3-18. http://www.cairn.info/revue-langue-francaise-2011-4-page-3.htm DOI : 10.3917/lf.172.0003

Goullier, François (2016). *Rapport de jury CAPES Externe, section Allemand, session 2016*. Ministère de l'Éducation Nationale, de l'Enseignement supérieur et de la Recherche. http://media.devenirenseignant.gouv.fr/file/externe/53/7/rj-2016-capes-externe-allemand_628537.pdf

Kilgarriff, Adam, Rychly, Pavel & Pomikalek, Jan (nd). *Sketch Engine*. Système de gestion de corpus. http://www.sketchengine.co.uk/

Kirchhoff, Frank & Primus, Beatrice (2016). Das Komma im mehrsprachigen Kontext [La virgule en contexte multilingue]. In Ralph Olsen, Christiane Hochstadt & Simona Colombo-Scheffold (dir.). *Ohne Punkt und Komma ...Beiträge zu Theorie, Empirie und Didaktik der Interpunktion*. Berlin : RabenStück Verlag, 78-97.

Landure, C. & Boulton, A. Corpus et autocorrection pour l'apprentissage des langues. *ASp* 57 | 2010, 11-30. http://asp.revues.org/931~;DOI:10.4000/asp.931

Leblay, Christophe (2009). En deçà du bien et du mal écrire. Pour une saisie en temps réel des invariants opérationnels de l'écriture *Pratiques* [En ligne], 143-144. http://pratiques.revues.org/1430 ; DOI : 10.4000/pratiques.1430

Leblay, Christophe (2014). Les écritures intermédiaires réflexives en littératie avancée. *Le Français Aujourd'hui – Pratiques de l'écrit en formation,* vol. 2014/1 (n° 184), 103-115.

O'Sullivan, Íde (2010). Using corpora to enhance learners' academic writing skills in French. *Revue française de linguistique appliquée* 2010/2 (Vol. XV), 21-35. https://www.cairn.info/revue-francaise-de-linguistique-appliquee-2010-2-page-21.htm

Schaeffer-Lacroix, Eva (2016). Impact de discussions métalinguistiques sur l'apprentissage de la production écrite en allemand, langue étrangère". In Sylvie Garnier, Fanny Rinck, Frédérique Sitri & Sarah de Vogüe (dir.). Linx (Revue des linguistes de l'université Paris Ouest Nanterre La Défense), vol. 72 : *Former à l'écrit universitaire, un terrain pour la linguistique ?*. Presses Universitaires de Paris Ouest, 193-211.

Trésor de la Langue Française Informatisé (nd). Dictionnaire en ligne. Nancy : Université de Lorraine. http://atilf.atilf.fr/

Verbes modaux et genres journalistiques : un éclairage statistique sur le français et l'allemand

Annalena Hütsch
Chaire de linguistique française, Université de Neuchâtel
annalena.hutsch@unine.ch

1. Introduction

Les verbes modaux du français et de l'allemand ont été analysés, d'un point de vue qualitatif, dans de nombreux travaux [1]. Notre communication propose d'examiner ces verbes (*devoir, falloir, pouvoir, vouloir ; dürfen, können, mögen, müssen, sollen, wollen*) dans une perspective différente [2]. À travers une étude de corpus, nous allons fournir des informations statistiques sur l'utilisation des verbes modaux dans la presse écrite française et allemande. Ce genre semble bien se prêter à une étude sur la modalité parce qu'il englobe une variété de sous-genres qui vont de la relation de l'information neutre et objective aux appréciations subjectives, avec par exemple les comptes rendus d'un côté et les éditoriaux de l'autre.

L'objectif de notre étude est de mettre en relief l'incidence du genre sur l'usage de la modalité verbale. Pour cela, nous mesurerons la fréquence et la spécificité d'apparition des verbes modaux pour chacune des deux langues dans un corpus comparable de presse et plusieurs sous-corpus qui correspondent aux sections des journaux retenus. Notre communication se divisera en deux grandes parties : (i) présentation des corpus et de la méthodologie utilisés, en spécifiant les défis que posent les plateformes d'interrogation pour les requêtes autour des verbes modaux ; (ii) présentation et interprétation des données quantitatives obtenues des corpus, en discutant les différences d'utilisation et les préférences d'apparition des verbes modaux en français et en allemand.

2. Corpus et méthodologie

Pour notre étude sur l'utilisation des verbes modaux dans la presse écrite et ses sous-genres, nous procéderons en plusieurs étapes. D'abord, nous interrogerons des corpus de nature différente – parallèle et comparable – à travers les bases *OPUS Word Alignment Database* (Tiedemann 2012) et *Emolex* (Diwersy *et al.* 2014) ; ensuite, nous appliquerons différents calculs – de fréquence et de spécificité – sur la base des données extraites des corpus.

Afin de constituer le paradigme des verbes modaux, nous utiliserons le corpus parallèle *Europarl3* (Koehn 2005). Cette collection recueille les discours prononcés dans le Parlement européen entre avril 1996 et octobre 2006 et elle contient 44 688872 tokens pour le français et 37 614344 tokens pour l'allemand. Nous interrogerons ce corpus à travers la base *OPUS Word Alignment Database* pour obtenir les équivalences entre verbes modaux français et allemands établies par alignement de mots automatique. Ainsi, en partant des verbes *devoir* et *pouvoir* en français (cf. Sueur 1979, Vetters 2004), puis de leurs équivalents allemands, nous retiendrons le paradigme mentionné ci-dessus.

Pour l'étude de la presse écrite en français et en allemand, nous utiliserons un corpus comparable issu de la base *Emolex* à travers la plateforme *BTLC* de l'Université de Cologne. Du côté de la presse française, nous retiendrons les journaux suivants pour notre corpus de travail : *Le Figaro* des années 2007 et 2008, *Le Monde* des années 2007 et 2008, ainsi que *Ouest France* des années 2007 et 2008 ; la taille du corpus est de 104 132 747 tokens. Du côté allemand, nous constituerons notre corpus journalistique de *Der Tagesspiegel* de l'année 2008, *Frankfurter Rundschau* des années 1999 et 2008, *Hamburger Abendblatt* de l'année 2008, ainsi que *Süddeutsche Zeitung* des années 2012 et 2013 ; la taille du corpus est de 229 098627 tokens. Même si les corpus français et allemand se distinguent par leur taille, ils sont comparables par leur genre, la période et la nature des textes, relevant de la presse quotidienne à la fois nationale et régionale des années 2000. Ce corpus de presse permet, par

1. Voir Diewald 1999, Öhlschläger 1989, Sueur 1979 et Vetters 2004, pour n'en citer que quelques-uns

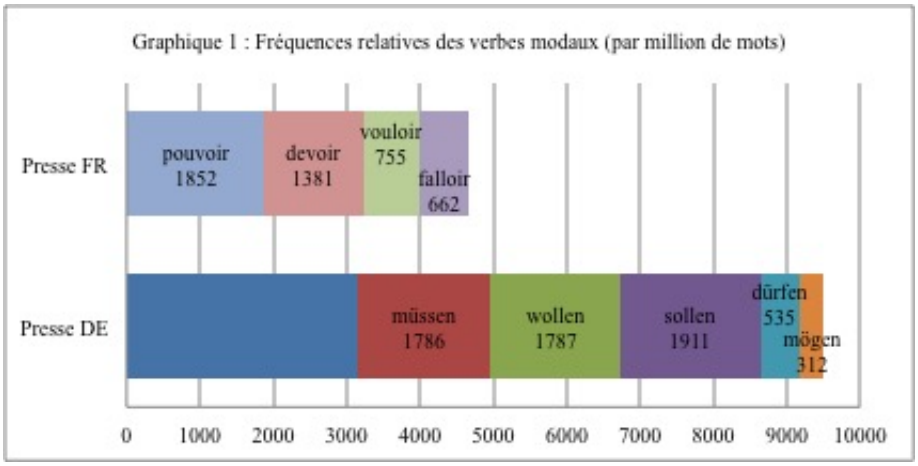
2. Notre étude s'inscrit dans le cadre d'un projet financé par le Fonds national suisse de la recherche scientifique intitulé « La représentation du sens modal et de ses tendances évolutives dans deux langues romanes : le français et l'italien» (no. $100012_{1}59458$), dont le cadre théorique et l'une des étapes de travail sont respectivement présentés dans Rossari (2016) et Rossari et al. (2016).

ailleurs, d'être partitionné en fonction de différentes sections thématiques ou génériques, qui peuvent se recouper d'un journal à l'autre.

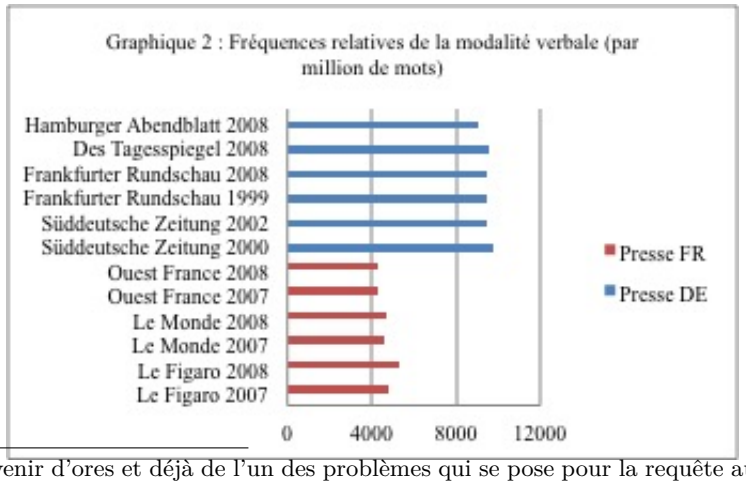
Dans un premier temps, nous calculerons des fréquences relatives pour déterminer le nombre d'occurrences des verbes modaux dans les corpus [3]. Cela nous permettra de comparer les proportions de la modalité verbale en français et en allemand. Dans un deuxième temps, nous calculerons des indices de spécificité pour déterminer dans quels types de journaux ou dans quelles sections les verbes modaux apparaissent en surnombre ou en sous-effectif. En d'autres termes, nous envisageons de révéler la représentativité des verbes modaux au sein de différents (sous-)corpus, afin de voir les préférences thématiques ou génériques de la modalité verbale en français et en allemand. Nous distinguerons les corpus globaux (journaux français vs journaux allemands) et différents sous-corpus pour chaque langue, constitués en fonction du journal (*Le Monde vs Le Figaro vs Ouest France*), en fonction de la distribution (journaux nationaux vs journaux régionaux) ou en fonction des sections (éditorial vs politique vs culture), afin de contraster l'usage des verbes modaux selon ces paramètres.

3. Données quantitatives

Les résultats des analyses quantitatives préliminaires permettent d'observer les tendances suivantes quant à l'usage des verbes modaux dans les corpus de presse.



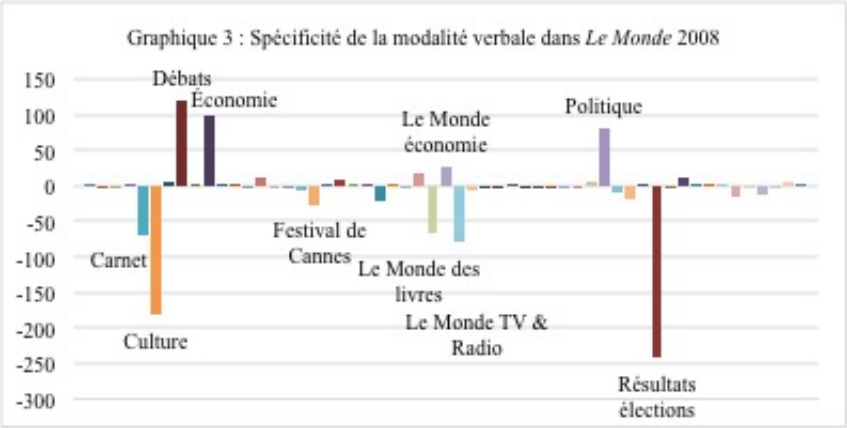
Tout d'abord, les fréquences relatives montrent une utilisation nettement plus importante des verbes modaux dans le corpus allemand par rapport au corpus français (cf. Graphique 1) [4] Par ailleurs, les fréquences relatives dans les différents journaux sont très similaires à l'intérieur de chaque corpus, français ou allemand, et semblent refléter une utilisation homogène de la modalité verbale dans la presse écrite française et allemande (cf. Graphique 2).

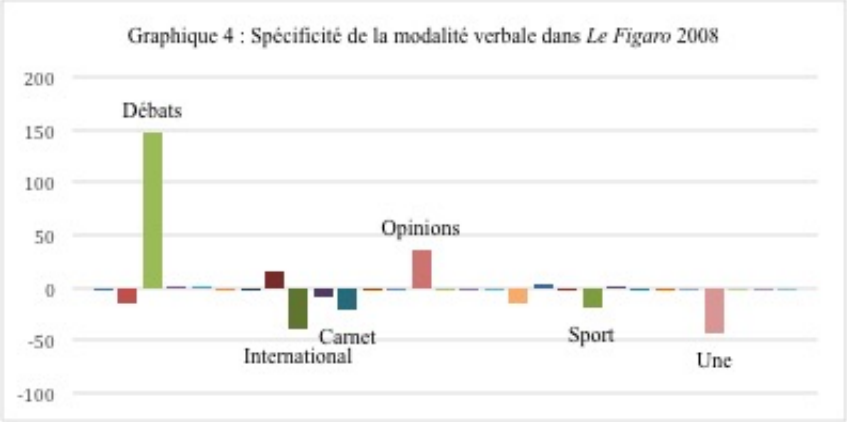


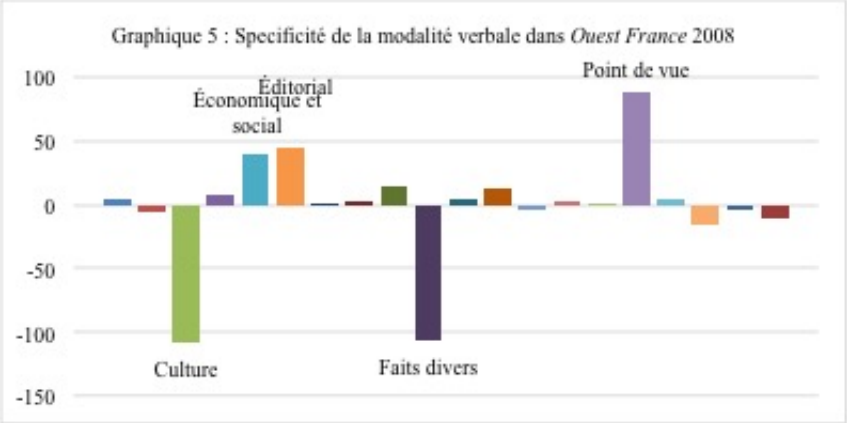
3. Nous tenons à prévenir d'ores et déjà de l'un des problèmes qui se pose pour la requête autour des verbes modaux, à savoir l'impossibilité d'écarter automatiquement l'utilisation de ces verbes dans leur fonction de verbe plein.

4. Le fait que le paradigme des verbes modaux soit plus important en allemand qu'en français (avec, pour la modalité du possible, *können*, *dürfen* et *mögen* du côté allemand vs *pouvoir* du côté français) ne semble pas pour autant affaiblir la tendance que le corpus allemand se montre « plus modal» quant à l'emploi de ces verbes.

Du côté français, les indices de spécificité quant aux occurrences des verbes modaux dans les différentes sections des journaux *Le Monde*, *Le Figaro* et *Ouest France* de l'année 2008 montrent des tendances similaires (cf. Graphiques 3 à 5) : l'usage de la modalité verbale apparaît être particulièrement spécifique des sections qui relèvent d'une expression subjective (débats, opinions, éditoriaux, points de vue) ou de prévisions (économie, politique), contrairement aux sections qui relatent des évènements (culture, carnet, sport, faits divers). Cette tendance reste à être confirmée du côté allemand. Par ailleurs, des analyses plus approfondies quant à la proportion des valeurs véhiculées par les verbes modaux dans quelques sections spécifiques pourraient fournir des informations supplémentaires sur l'utilisation des verbes modaux dans la presse écrite française et allemande.







Références bibliographiques

Diewald, G. (1999). *Die Modalverben im Deutschen : Grammatikalisierung und Polyfunktionalität*. Berlin : de Gruyter.

Diwersy, S., Goossens, V., Grutschus, A., Kern, B., Kraif, O., Melnikova, E., Novakova, I. (2014). Traitement des lexies d'émotion dans les corpus et les applications d'EmoBase. *Corpus, 13*, 269-293.

Koehn, P. (2005). Europarl : A Parallel Corpus for Statistical Machine Translation. *MT Summit*, 79-86.

Rossari, C. (2016). Les fluctuations de sens dans quelques formes modales à la lumière d'une approche quantitative et qualitative. *Studii de lingvistică*, 6, 127-142.

Rossari, C., Hütsch, A., Ricci, C., Salsmann, M. et Wandel, D. (2016). Le pouvoir attracteur de mais sur le paradigme des adverbes épistémiques : du quantitatif au qualitatif. *Proceedings of 13th International Conference on Statistical Analysis of Textual Data*, II, 819-823.

Öhlschläger, G. (1989). *Zur Syntax und Semantik der Modalverben des Deutschen.* Tübingen : Niemeyer.

Sueur, J.-P. (1979). Une analyse sémantique des verbes *devoir* et *pouvoir*. *Le français moderne, 47(2)*, 97-120.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation*, 2214-2218.

Vetters C. (2004). Les verbes modaux pouvoir et devoir en français. *Revue belge de philologie et d'histoire, 82(3)*, 657-671.

Session 6.A.
Sémantique

Étude de l'évolution sémantique des prépositions *à*, *en*, *dans*, *dedans* du français. Quel(s) apport(s) d'une périodisation automatique ?

Sascha Diwersy [1], Achille Falaise [2] et Denis Vigier [3]
[1] Montpellier 3, UMR Praxiling
[2] ENS Lyon, UMR ICAR
[3] Lyon 2, UMR ICAR
sascha.diwersy@univ-montp3.fr, achille.falaise@ens-lyon.fr, denis.vigier"@univ-lyon2.fr

Le corpus diachronique du français PRESTO, constitué dans le cadre du projet franco-allemand du même nom (http ://presto.ens-lyon.fr), échantillonné, contrôlé et équilibré (critères chronologiques, d'auteurs et de genres discursifs), permet aujourd'hui d'étudier en diachronie longue (1509-1944) l'évolution du profil cooccurrentiel des prépositions du français. C'est ce que nous proposons ici de faire pour les prépositions *à, en, dans, dedans*, en nous posant plus spécifiquement la question de la périodisation des phénomènes cooccurrentiels sur lesquels nous travaillons.

Rappelons d'abord qu'on peut tenir pour acquis (voir les nombreuses études sur corpus qui toutes s'accordent sur ce point : Darmesteter 1885 ; Gougenheim 1945, 1951 ; Brunot 1967 ; Fagard & Sarda 2009 ; etc.) que c'est à partir du début de la seconde moitié du XVIe s. que *dans* fait une entrée remarquable sur la scène des prépositions du français – du moins dans le discours littéraire pour figurer finalement, en français contemporain, au cinquième rang en termes de fréquence d'emploi derrière *de, à, en* et *pour*[1]. Or si l'on examine au moyen du calcul des cooccurrences sur TXM (calcul des spécificités de P. Lafon) et sur PrimeStat (calcul du log likelihood) quels sont les accompagnateurs nominaux préférés de dans lors de ses premiers emplois, on observe que les noms qui figurent parmi ses collocatifs les plus spécifiques (au sens statistique) dénotent tous une réalité dotée d'une extension matérielle ou physique : lieux géographiques construits de la vie quotidienne (*château, ville, maison, salon, ...*) ou naturels (*mer, bois, plaine, ...*), noms désignant des parties et des productions du corps humain (*entrailles, sang, sein*). Ces réalités constituent, en discours, des sites de repérage pour une cible, au terme (1) ou non (2) d'un déplacement dans l'espace :

1. *Le president entra dans la chambre et trouva sa femme et nicolas couchez ensemble.* (1550, M. de Navarre, *L'Heptameron*)

2. *Ce pendant qu'on est en ce monde, On est dans une mer profonde (...)* (1587, Pierre de l'Estoile, *Registre-journal du regne de Henri III*)

Ce profil cooccurrentiel de *dans* qui s'affirme lors de ses premiers emplois en français préclassique fait de cette préposition nouvellement arrivée **une rivale distributionnelle** de trois autres prépositions beaucoup plus anciennes :

- *en* qui, issue du latin, couvre un ensemble d'emploi qui se partagent de manière équilibrée entre régimes nominaux déterminés et non déterminés et qui s'avère apte elle aussi à opérer des localisations de cible dans les limites d'un site ; par ex.

    3. *Quant le mary veid qu'il en avoit bien faict son debvoir, entra en la chambre et le mercia de la peyne qu'il en avoit prinse* (1550, M. de Navarre, *l'Heptameron*)
    4. *Mais tousjours demouroit en la nef entre les femmes* (1532, F. Rabelais,*Pantagruel*)

- *dedans*, qui semble se distinguer de en par sa préférence pour des contextes où la cible est entièrement recouverte par le site :

    5. *Or on ne peut cognoistre cela aux poissons, car on ne peut sçavoir leur aage, d'autant qu'ils vivent dedans l'eau.* (1556, B.-G. Gelli, *Les discours fantastiques de Julien Tonnelier*)

1. Nous avons étudié ailleurs les raisons de cette surprenante fortune de *dans* en français.

- Quant à *à*, on sait que les formes amalgamée *au, aux* qui en étaient issues pouvaient être recrutées aussi bien dans les contextes où cette préposition se trouvait au contact des formes *le, les* de l'article défini, que dans les contextes où les formes issues de la fusion de *en* et de *le/les*[2] étaient déficientes – *au, aux* entrant alors dans le paradigme de *en* :

    6. *Face le ciel (quand il voudra) revivre / Lisippe, Apelle, Homere, qui le pris / Ont emporté sur tous humains espris / En la statue, au* [= **en* le] tableau, *et au* [= **en* le] *livre.* (1550, Du Bellay, *L'Olive*) (cité par S. Lardon & M.-C Thomine, *ibid.* : 393)

En d'autres termes, tout porte à croire que les « destins» cooccurrentiels de *à, en, dans, dedans* ont été – du moins pendant une certaine période après 1550 – peu ou prou liés. Nous nous proposons donc dans cette communication d'étudier **l'évolution sémantique de ces quatre prépositions entre le XVI e s. et le XX e s.**, notre hypothèse étant que l'irruption de *dans* sur la scène des prépositions a profondément recomposé le profil distributionnel des trois autres.

Après avoir brossé le contexte au sein duquel *dans* paraît vers 1550, nous présenterons les résultats d'une étude comparée que nous avons accomplie sur le corpus PRESTO 1501-1950 et portant de l'évolution des profils cooccurrentiels de ces quatre prépositions. Cette étude s'adosse (i) d'une part à une série de calculs de spécificités conduits sur une partition du corpus diachronique PRESTO (tranches de cinquante années découpées arbitrairement en s'appuyant sur les frontières de (demi-)siècles), (ii) d'autre part sur une suite de calculs de cooccurrence lexicale conduits sur des sous-corpus et prenant pour pivot successivement *dans, en, dedans, à*. Les résultats de ces calculs ont été analysés manuellement, en vue d'identifier pour chacune des prépositions d'éventuelles zones temporelles où apparaissent des recompositions particulièrement prononcées de leur profil combinatoire.

Puis nous nous tournerons vers la méthode de *classification ascendante hiérarchique par contiguïtés* (CAHC, de l'anglais *Variability based neighbour clustering* - cf. Gries & Hilpert 2008 ; 2012 ; Hilpert 2013 : 32-45), qui vise avant tout à produire une périodisation automatique. Cette méthode peut par exemple s'appliquer aux paradigmes cooccurrentiels de *en*, définis en termes des collocatifs nominaux spécifiques[3] relevés, dans une fenêtre de trois mots à droite, et par tranches de 25 années, dans le corpus PRESTO étendu (1501-2010). On peut alors constater (figure 1), que les emplois de cette préposition font l'objet d'une reconfiguration fondamentale, qui est entamée pendant la première moitié du XVIIe siècle et qui va depuis dans le sens d'une spécialisation (cf. Diwersy, Falaise, Lay et Souvay 2017).

2. El, on, ou, . . . / els, es, ès, ez, . . .

3. Ont été retenus comme spécifiques tous les collocatifs nominaux dont le score d'association *log-likelihood* était égal ou supérieur à 10,83 (valeur-p $< 0{,}001$)

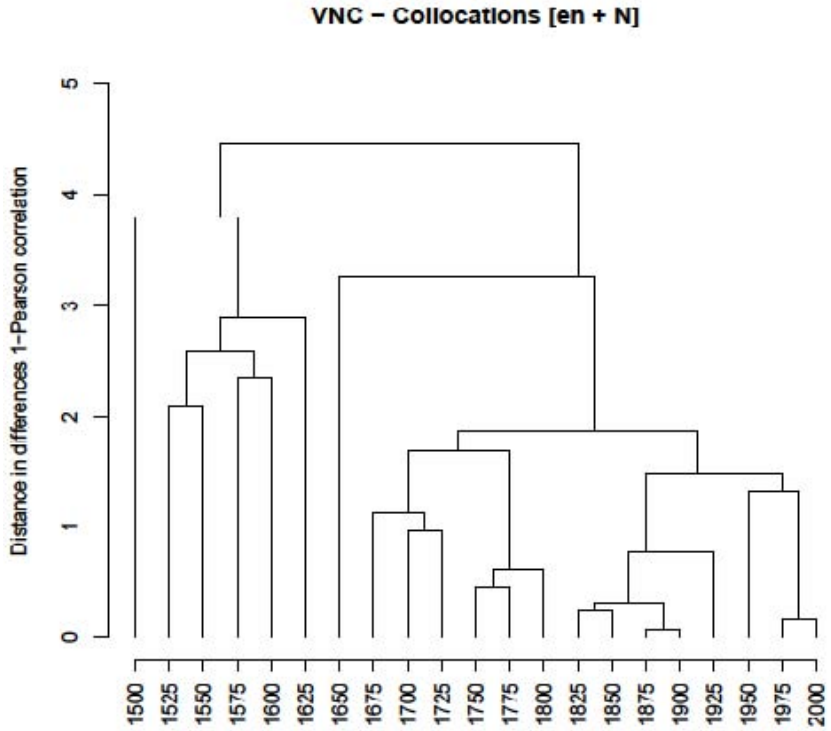


FIGURE 1 – CAHC appliquée aux paradigmes collocationnels formés par les noms constituant un SN régi par en (corpus PRESTO étendu, tranches de 25 années).

Notre objectif sera de comparer les résultats obtenus par les deux méthodes d'analyse ci-dessus, la première, « hybride » (« analyse des collocations », au sens de K. Heylen et A. Bertels (2016 : 53)), en ceci qu'elle combine identification automatique et statistique des indices contextuels et traitement manuel des résultats, et la seconde, entièrement automatique (méthodes d'« analyse distributionnelles»).

Références bibliographiques

BRUNOT F. (1967), *Histoire de la langue française. Des origines jusqu'en 1900.* Tome II : *le seizième siècle*, Paris : A. Colin.

DARMESTETER A. (1885), *Notes sur l'histoire des prépositions françaises en, enz, dedans, dans*, Paris : Le Cerf.

DIWERSY S., FALAISE A., LAY M.-H., SOUVAY G. (2017). Ressources et méthodes pour l'analyse diachronique. *Langages* 206, à paraître.

FAGARD B. & SARDA L. (2009), « Etude diachronique de la préposition *dans* ». *in* J. François, E. Gilbert, C. Guimier et M. Krause (dir.), *Autour de la préposition : position, valeurs, statut et catégories apparentées à travers les langues*, Bibliothèque de Syntaxe et Sémantique, Caen : PUC, 225-236.

GOUGENHEIM G. (1945), « Les prépositions « en » et « dans » dans les premières oeuvres de Ronsard ». *in Etudes de grammaire et de vocabulaire français, réunies sur l'initiative de ses collègues et amis pour son soixante-dixième anniversaire*, Paris : Picard, 55-65.

GOUGENHEIM G. (1950, 1970), « Valeur fonctionnelle et valeur intrinsèque de la préposition « en » en français moderne », in *Etudes de grammaire et de vocabulaire français, réunies sur l'initiative de ses collègues et amis pour son soixante-dixième anniversaire*, Paris : Picard, 55-65.

GRIES S.T. & HILPERT M. (2008), « The identification of stages in diachronic data : variability-based neighbour clustering», *Corpora* 3 (1), 59-81.

GRIES S.T. & HILPERT M. (2012), « Variability-based neighbor clustering : a bottom-up approach to periodization *in* historical linguistics », in T. Nevalainen & E. Traugott (eds), *The Oxford handbook of the history of English*, Oxford : Oxford University Press, 134-144.

HEYLEN, K., BERTELS, A. (2016), « Sémantique distributionnelle en linguistique de corpus», *Langages*, 201, 51-64

HILPERT M. (2013), *Constructional Change in English : Developments in Allomorphy, Word formation, and Syntax*, Cambridge : Cambridge University Press.

Proximité sémantique des dérivés morphologiques

Marine Wauquier [1], Cécile Fabre [2] et Nabil Hathout [2]
[1] Université de Toulouse, France
[2] CLLE-ERSS, CNRS & Université de Toulouse
marine.wauquier@etu.univ-tlse2.fr, cfabre@univ-tlse2.fr, hathout@univ-tlse2.fr

1. Objectifs de l'étude

Dans cette étude, nous nous intéressons à la proximité sémantique entre les lexèmes qui appartiennent à une même famille dérivationnelle, en considérant plus particulièrement les verbes et leurs noms d'agent et d'action dérivés. Le point de départ de ce travail est le souhait de tester à l'aide de critères objectifs l'hypothèse de Roché (2009) selon laquelle le sens d'une base (ex : *protéger*) et de son nom d'action dérivé (ex : *protection*) seraient identiques : la variation formelle et catégorielle, simple conséquence d'un choix de construction syntaxique, ne s'accompagnerait pas d'une variation sémantique significative. Nous cherchons à vérifier si cette hypothèse se traduit en termes de profil distributionnel des lexèmes en corpus, et ce de façon homogène. Nous considérons les liens de proximité sémantique au sein des triplets (verbe, nom d'agent, nom d'action) comme (*poncer, ponceur, ponçage*). L'hypothèse est que le verbe (*poncer*) est plus proche sur le plan distributionnel de son nom d'action (*ponçage*) qu'il ne l'est de son nom d'agent (*ponceur*). Ce travail est réalisé à partir d'une ressource dérivationnelle, *Lexeur*, comportant 5 974 familles dérivationnelles, projetée sur 2 corpus tirés de *Wikipedia* et du *Monde*.

La proximité sémantique entre le verbe et ses dérivés a fait l'objet de nombreux travaux, principalement focalisés sur la prise en compte de deux types de critères : la préservation de la structure argumentale du verbe (Grimshaw 1990) et l'héritage de propriétés sémantiques, en particulier aspectuelles, du verbe par le nom (Haas et al. 2013). Ces travaux se fondent généralement sur l'application de tests d'acceptabilité, éventuellement complétés par des procédures d'annotation de corpus (Balvet et al. 2011). Nous avons fait le choix d'une autre approche, visant à considérer l'usage de ces lexèmes en corpus, en nous appuyant sur le critère de proximité distributionnelle comme indice de proximité sémantique. Les méthodes automatiques fondées sur une approche distributionnelle du sens connaissent aujourd'hui un succès important en traitement automatique des langues (Sahlgren 2010, Fabre et Lenci 2015). Les linguistes ont désormais à leur disposition des outils de calcul qui mettent en œuvre à très large échelle, sur de vastes corpus, l'hypothèse harrissienne selon laquelle la proximité sémantique entre les mots peut être assimilée à leur degré de proximité distributionnelle. Ces outils ont été récemment intégrés dans l'examen de questions linguistiques variées, en particulier en morphologie (Lazaridou et al. 2013, Zeller et al. 2014, Kisselew et al. 2015).

L'approche distributionnelle nous permet de tester à grande échelle l'hypothèse de Roché et de dégager des tendances générales afin de répondre à la question suivante : au sein du triplet (verbe, nom d'agent, nom d'action), la paire Verbe – Nom d'action présente-t-elle systématiquement le degré de proximité distributionnelle le plus élevé ? Dans un deuxième temps nous nous consacrons à un examen plus détaillé, appuyé sur des observations en corpus, de types de familles dérivationnelles exhibant des propriétés distributionnelles particulières.

2. Données de l'étude : corpus et ressource dérivationnelle Lexeur

L'utilisation d'un outil de calcul distributionnel automatique requiert l'analyse de corpus de grande taille. Nous avons opté pour l'utilisation de deux corpus : le corpus *Wikipedia* est issu de la version française de 2013 de l'encyclopédie en ligne. Il compte près de 255 millions de mots. Ce choix est guidé par le souhait de disposer d'un vocabulaire vaste et varié, relevant de domaines hétérogènes, à l'image de la diversité des lexèmes que nous étudions. Le corpus *LM10* est composé des articles du journal *Le Monde* publiés entre les années 1991 et 2000 et contient près de 200 millions de mots. Nous l'utilisons pour tester la stabilité des observations réalisées sur *Wikipedia*.

La ressource Lexeur [1] est un lexique dérivationnel qui regroupe 5974 noms en *–eur*. Pour simplifier,

1. La constitution de la ressource *Lexeur* a été réalisée au sein de l'équipe CLLE-ERSS, sous la coordination de Cécile Fabre et Nabil Hathout. Ce projet a bénéficié en 2001 d'un financement ILF.

Nom d'agent masc.	Nom d'agent fém.	Base	Cat.	Autres dérivés
abatteur/Ncms	abatteuse/Ncfs	abattre/Vmn—-	Vb	abat/Ncms ; abattement/Ncms abatture/Ncfs ; abattage/Ncms abattis/Ncms
endoscopeur/Ncms	endoscopeuse/Ncfs	∅		endoscopie/Ncfs
fraudeur/Ncms	fraudeuse/Ncfs	frauder/Vmn—-	Vb	fraude/Ncfs
wheeleur/Ncms	wheeleuse/Ncms	wheel/Ncms	Nb	∅

TABLE 1 – extrait de Lexeur

Nom d'agent	Base verbale	Nom d'action	P(VbAg)	P(AgAc)	P(VbAc)
discriminateur	*discriminer*	*discrimination*	0,15	0,15	0,49
directeur	*diriger*	*direction*	0,44	0,52	0,27

TABLE 2 – exemples de triplets valués

nous parlons de noms d'agent, mais *Lexeur* regroupe indistinctement des noms d'agent et des noms d'instrument. Ces noms sont issus du *Trésor de la Langue Française*, complétés par des attestations issues du Web. Comme illustré dans le tableau 1, le nom en -eur a été associé, par une procédure d'annotation manuelle, à sa famille constructionnelle, composée d'un nom d'agent féminin, de la base (verbale ou nominale), et d'une liste de tous les noms processifs identifiés. Chaque lexème reçoit une étiquette morphosyntaxique.

Les exemples présentés dans le tableau 1 montrent la diversité des familles constructionnelles : certaines sont très fournies, comme dans le cas d'*abatteur*, d'autres peuvent être lacunaires comme celle d'*endoscopeur* (sans base verbale identifiée) ou de *wheeleur* (qui a seulement un dérivé agentif).

Nous extrayons des entrées de Lexeur l'ensemble des triplets de la forme « nom d'agent en *–eur* – base verbale – nom d'action». Les noms d'action sont extraits de la colonne « autres dérivés » illustrée dans le tableau 1, sans tenir compte de la polysémie du lexème dont les interprétations en contexte peuvent varier (Huygue 2014). On dispose à ce stade de 13136 triplets.

3. Méthode d'analyse distributionnelle

Nous utilisons Word2Vec (Mikolov et al. 2013), un outil d'apprentissage non supervisé qui fournit une représentation vectorielle du sens des mots d'un corpus et exploite cette représentation à l'aide de différents modules calculant les voisins distributionnels des mots ou le score de proximité distributionnelle entre plusieurs mots. Nous avons utilisé les paramètres par défaut de l'outil [2] pour construire la matrice distributionnelle des 2 corpus. Le calcul distributionnel est donc basé sur l'examen de cooccurrences lexicales dans une fenêtre contextuelle donnée, sans prise en compte d'informations syntaxiques. Word2Vec attribue un score de proximité distributionnelle (P) aux 3 couples de lexèmes qui composent les triplets extraits à l'étape précédente (tableau 2). Ces couples sont notés : VbAg (verbe – nom d'agent) ; AgAc (nom d'agent – nom d'action) ; VbAc (verbe – nom d'action). Le score de proximité varie de 0 (proximité nulle) à 1 (proximité maximale pour 2 formes dont les propriétés distributionnelles sont identiques).

Le tableau 2 montre deux triplets au comportement distributionnel différent : dans le cas de *discriminateur*, la proximité du verbe et du nom d'action est nettement supérieure à celle qui caractérise les deux autres couples de lexèmes. Dans le cas de *directeur*, le couple VbAc est au contraire le moins proche distributionnellement.

1945 triplets valués ont été extraits du corpus *Wikipedia*, et 1520 du corpus *LM10*. Cette forte réduction du nombre de triplets par rapport à l'ensemble issu de Lexeur s'explique par le filtre opéré par le corpus traité par Word2Vec : seuls sont retenus les lexèmes dépassant le seuil de fréquence requis dans les paramètres de l'outil.

2. Word2Vec utilise par défaut l'architecture CBOW, l'algorithme d'entraînement Negative Sampling, un seuil minimum de fréquence de 5, un seuil de sous-échantillonnage des mots fréquents de 1-3, une taille de fenêtre de 5, et comme nombre de dimensions des vecteurs 100.

4. Premiers résultats

Un premier niveau de résultats permet de dégager des tendances sur les 2 corpus. Ainsi, sur les 1945 triplets observés dans Wikipedia, 58% présentent un score de proximité distributionnelle plus élevé dans le cas du couple VbAc. Dans 25% des cas (488 triplets), c'est le couple AgAc qui domine, et dans 17% des cas (324 triplets), le couple AgVb. On retrouve cette tendance, mais avec des nuances importantes, pour le corpus LM10 (respectivement 50% ; 31% ; 19%). Le score moyen de proximité dans *Wikipedia* pour le couple VbAc est de 0,39, pour le couple AgAc de 0,29 et pour le couple AgVb de 0,25. Le score moyen de proximité pour chaque couple dans le corpus LM10 est du même ordre, quoique légèrement inférieur (respectivement 0,34 ; 0,28 ; 0,25). Ces premiers résultats vont dans le sens de l'hypothèse de Roché : au sein des 3 combinaisons examinées, c'est le couple verbe – nom d'action qui exhibe le score de proximité le plus élevé. Mais ce n'est qu'une tendance et dans de nombreux cas les usages en corpus s'écartent de ce principe. Le fait de disposer d'un score de proximité distributionnelle nous permet de dégager différents cas de figure.

Dans une deuxième étape, nous utilisons ce score pour mener des analyses plus approfondies. Nous nous intéressons au cas des triplets qui exhibent un score de proximité très faible. Cette information permet de mettre en évidence des cas d'association douteuse et amène à interroger la validité de leur recensement dans Lexeur (ex : *ouvreuse – ouvrer – ouvrage*, dont les 3 relations ont un score cumulé de 0,08), ou à repérer des usages spécifiques en corpus. Nous examinons également le comportement distributionnel des noms en fonction de l'opération dérivationnelle (suffixation en *-eur*, *-euse*, *-rice* pour les noms d'agent ; conversion ou suffixation en *-age*, *-ment*, *-tion*, *-ure*, etc. pour les noms processifs), de manière à différencier des degrés de proximité en fonction de cette opération.

Références bibliographiques

Balvet, A., Barque, L., Condette, M. H., Haas, P., Huyghe, R., Marin, R., & Merlo, A. (2011). La ressource Nomage. Confronter les attentes théoriques aux observations du comportement linguistique des nominalisations en corpus. *Traitement Automatique des Langues*, 52(3), 129-152.

Fabre, C., et Lenci, A. (2015). Distributional Semantics Today – Introduction to the special issue. *Traitement Automatique des Langues*, 56(2), 7-20.

Grimshaw, J. (1990). *Argument structure*. the MIT Press.

Haas, P., Huyghe, R., & Marin, R. (2008). Du verbe au nom : calques et décalages aspectuels. In *Congrès Mondial de Linguistique Française*, Paris, 2051-2065.

Huyghe, R. (2014). La sémantique des noms d'action : quelques repères. *Cahiers de lexicologie*, 105, 181-201.

Kisselew, M., S. Padó, A. Palmer, and J. Šnajder (2015). Obtaining a better understanding of distributional models of german derivational morphology. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015)*, London.

Lazaridou, A., Marelli, M., Zamparelli, R. et Baroni, M. (2013). Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria.

Mikolov, T., Chen, K., Corrado, G., et Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of International Conference on Learning Representations* (ICLR).

Roché, M. (2009). Pour une morphologie lexicale. *Mémoires de la Société de Linguistique de Paris*, (Nouvelle serie n°17), 65-87.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, 20(1), 33-54.

Zeller, B. D., S. Padó, and J. Snajder (2014). Towards semantic validation of a derivational lexicon. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics* : Technical Papers (COLING 2014), Dublin, pp. 1728–1739.

Les sens de *numérique* – émergence d'emplois et dynamique du changement sémantique

Sascha Diwersy , Francesca Frontini , Agata Jackiewicz , Giancarlo Luxardo et Agnès Steuckardt
Praxiling UMR 5267, CNRS - Université Paul-Valéry Montpellier 3
prenom.nom@univ-montp3.fr

1. Introduction

L'avènement des nouvelles technologies de l'information et de la communication s'accompagne de modifications substantielles de nos modes de vie et de nos manières de penser et de parler. Dorénavant, nous vivons à l'ère numérique, dans une société numérique. Nous manipulons des outils numériques avec plus ou moins de « bons sens numérique », bénéficions des services numériques dans des administrations, feuilletons des livres numériques, pratiquons une pédagogie numérique, participons à une démocratie numérique, via des Agoras numériques... La question du numérique se trouve ainsi au cœur d'enjeux sociétaux présentant des aspects interdépendants et parfois contradictoires. Les usages se multiplient, alors que les emplois établis (en rapport avec les nombres tout court) persistent.

Notre contribution s'intéresse à l'évolution diachronique des emplois de l'adjectif *numérique* en français contemporain. Elle est structurée de la façon suivante : la première partie est consacrée à la présentation du corpus utilisé dans notre étude et d'une méthodologie originale dédiée à la classification de corpus diachroniques. Dans la deuxième partie, nous exposons les résultats obtenus grâce aux traitements lexicométriques effectués pour proposer ensuite, dans la dernière partie, un bilan interprétatif et quelques perspectives de recherche.

2. Corpus et méthodologie d'exploitation lexicométrique

2.1. Corpus

Le corpus sur lequel est basée notre étude est composé d'articles de presse parus dans le quotidien *Le Monde* entre 1945 et 2015. Pour constituer cet échantillon, nous avons recueillli entre 70 et 80 numéros par année (soit de 1 à 2 numéros par semaine), pour un total de 671.943 textes et de 348.497.000 mots-occurrences. Le corpus a fait l'objet d'un étiquetage en parties du discours, d'une lemmatisation et d'une annotation en relations de dépendance syntaxique au moyen de la chaîne de traitement Bonsaï (Candito *et al.* 2010) dans sa version qui fait appel à l'analyseur MaltParser (Nivre *et al.* 2006).

2.2. Méthodologie d'exploitation lexicométrique

Notre étude de l'évolution du sémantisme propre au lexème *numérique* est basée sur l'observation de la variation diachronique de son profil combinatoire, tel qu'il se manifeste à travers les collocations adjectivo-nominales impliquant *numérique* en tant qu'épithète. Nous avons donc partitionné le corpus en tranches de cinq ans (1945-1949, 1950-1954, 1955-1959, etc.), pour lesquelles nous avons respectivement établi un inventaire cooccurrentiel des noms régissant l'adjectif *numérique*. Ces inventaires ont été répertoriés sous forme d'une liste, que nous avons soumise à un calcul de scores d'association, effectué dans la lignée de la méthodologie relevant de l'analyse dite collostructionnelle (Stefanowitsch & Gries 2003), au moyen du test exact de Fisher-Yates. Les résultats obtenus ont été enregistrés sous forme d'un lexicogramme (Heiden & Tournier 1998), dont le tableau suivant donne quelques extraits à titre illustratif :

Pour étayer notre analyse de l'évolution diachronique des emplois de l'adjectif *numérique*, nous avons appliqué à ce lexicogramme un procédé de périodisation automatique, à savoir la Classification Ascendante Hiérarchique par Contiguïtés (CAHC), dérivé de la méthode de classification proposée par Gries & Hilpert (2008 ; 2012) : Variability-based Neighbor Clustering (VNC). Cette méthode a été conçue pour le traitement de données lexico-statistiques relevant d'une variable d'ordre chronologique. La classification que nous avons mise en œuvre vise l'évolution des (scores de) dissimilarités entre les inventaires cooccurrentiels adjacents au sein de la série constituée par les tranches temporelles issues

Mot-pôle	Collocatif	Sous-échantillon / Partie (tranche de 5 ans)	Co-fréquence	Score Fisher-Yates
numérique	importance	1945-1949	17	51,4177
...	...	...	...	...
numérique	supériorité	1950-1954	19	72,1411
...	...	...	...	...
numérique	commande	1975-1979	17	51,6906
...	...	...	...	...
numérique	télévision	1990-1994	38	84,7942
...	...	...	...	...
numérique	révolution	2010-2015	95	191,6682
...	...	...	...	...

TABLE 1 – Extrait du lexicogramme répertoriant les collocations adjectivo-nominales impliquant *numérique* en tant qu'épithète

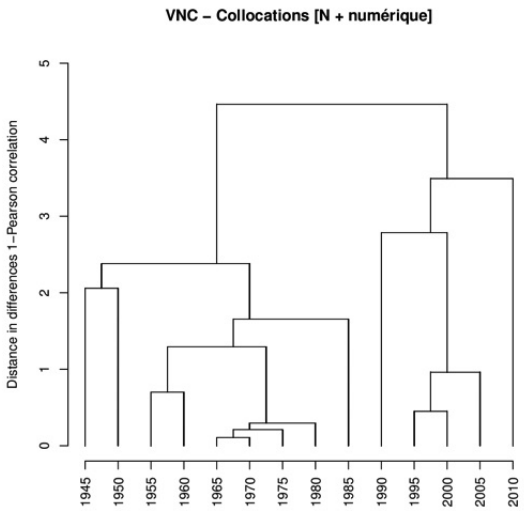


FIGURE 1 – CAHC appliquée aux collocations adjectivo-nominales impliquant *numérique* en tant qu'épithète

du partitionnement du corpus et par conséquent à mettre en évidence les discontinuités majeures que nous avons effectué en amont de l'analyse cooccurrentielle.

3. L'évolution diachronique des emplois de *numérique*

La CAHC appliquée au lexicogramme de *numérique* permet de produire un schéma combinant l'évolution dans le temps avec une classification hiérarchique (voir ci-dessous). Celle-ci met en évidence deux « coupes » majeures et trois périodes qui se dégagent successivement : 1945-1989, 1990-2009, 2010-2015 (le premier basculement illustré par le schéma intervenant sur 1990, le deuxième sur 2010).

Ces deux coupes peuvent être expliquées par l'extraction des collocatifs nominaux de la première tranche de la période, marquant la rupture avec la période précédente. En se limitant au premier décile (premiers 10

- pour la période 1990-1994 (marquant la rupture avec les périodes précédentes) :

  télévision, compression, image, commande, TVHD, technologie, norme, définition, transmission, donnée, supériorité, infériorité

- pour la période 2010-2015 :

révolution, livre, économie, tablette, télévision, ère, technologie, plate-forme, version, radio, support, outil, art, fichier, photo, transformation, oubli, simulation, bouquet, écosystème, baladeur, caméra, image, infériorité, donnée, modèle, usages, contenu, univers, plateforme, bibliothèque, développement, mutation, édition, abonné, format, Interview, identité, média, Métropole, lecture, stratégie, offre, création, service, effet, trace, Syntec, information, appareil, innovation, culture

Si on part des quatre emplois majeurs [1] indiqués par *Le Petit Robert* 2016, on voit que les deux premières coupes signalées par la CAHC sont induites avant tout par des collocations impliquant l'emploi *numérique4* (cf. *Le Petit Robert* 2016, s.v. NUMERIQUE : « Se dit de la représentation de données, de grandeurs physiques au moyen de caractères tels que des chiffres (opposé à *analogique*), ainsi que des procédés utilisant ce mode de représentation (. . .) »), ce qui permet de dresser un bilan intermédiaire, qui consiste à constater d'une part, par rapport à la période 1990-1994, un « enracinement » évident de l'emploi *numérique4* renvoyant, par oppositon à *analogique*, à une technologie et aux appareils qui utilisent celle-ci et d'autre part, pour la période à partir de 2010, une forte présence de collocations marquant différentes extensions métonymiques de cet emploi, dont la plus emblématique concerne la qualification des domaines et phénomènes sociétaux (*économie, art, univers, écosystème – révolution, ère, développement, mutation*) relatifs à cette technologie, à son utilisation et à ses produits. Le changement de collocation repéré dans l'usage de l'adjectif *numérique* à partir de 2010 correspond sans doute à une progression des usages sociaux d'internet. Il semble que cet emploi, qui apparaît, du point de vue du parcours sémantique diachronique et logique comme un emploi plutôt restreint, tende à devenir le sens prototypique de *numérique*. L'usage actuel de cet adjectif se caractérise à la fois par ce caractère massif et par ce brouillage sémantique.

4. Perspectives de recherche

L'évolution des emplois de l'adjectif *numérique* telle qu'on peut l'observer grâce à l'analyse collocationnelle que nous venons de présenter amène plusieurs questionnements.

D'un point de vue sémantique, en étudiant les classes de collocatifs nominaux, il y a lieu de se demander si l'évolution récente des emplois qui se manifeste à travers ces classes ne s'ouvre pas à des emplois subjectifs et en particulier axiologiques, comme le montrent des collocatifs tels que *fracture*, *vulnérabilité* ou encore *innovation*. A cet égard, il peut s'avérer très instructif d'étudier les emplois en extension de numérique en observant son fonctionnement dans le cadre de constructions relevant d'une grammaire locale telles que « N numérique de GN » (traces numériques des usagers, vulnérabilité numérique des voitures connectées, souveraineté numérique des pays, sabotage numérique des installations industrielles . . .).D'un point de vue sémantique, en étudiant les classes de collocatifs nominaux, il y a lieu de se demander si l'évolution récente des emplois qui se manifeste à travers ces classes ne s'ouvre pas à des emplois subjectifs et en particulier axiologiques, comme le montrent des collocatifs tels que fracture, vulnérabilité ou encore innovation. A cet égard, il peut s'avérer très instructif d'étudier les emplois en extension de numérique en observant son fonctionnement dans le cadre de constructions relevant d'une grammaire locale telles que « N *numérique* de GN » (*traces numériques des usagers*, *vulnérabilité numérique des voitures connectées, souveraineté numérique des pays, sabotage numérique des installations industrielles* . . .).

D'un point de syntaxique et morphologique, il semble pertinent de se poser la question de savoir si la dynamique du changement sémantique qu'on peut constater par rapport à l'extension des emplois de *numérique* va de concert avec des changements aux niveaux syntaxique (cf. la conversion vers l'emploi nominal de *numérique* adjectif) ou dérivationnel.

Au-delà de ces questions, d'autres explorations pourraient compléter le programme de travail que nous venons d'exposer, et notamment la prise en compte de productions langagières plus actuelles et de différents genres et registres (presse *vs* messages de forum *vs* textes scientifiques, etc.). De même, il sera indispensable d'entrependre une étude contrastant *numérique* avec *digital*, *informatique* ou *télématique*.

1. Les emplois majeurs de l'adjectif *numérique* selon *Le Petit Robert* 2016 sont : *numérique1* : « Qui est représenté par un nombre, se fait avec des nombres » ; *numérique2* : « Qui concerne les nombres arithmétiques » ; *numérique3* : « Évalué en nombre » ; *numérique4* : « Se dit de la représentation de données, de grandeurs physiques au moyen de caractères tels que des chiffres (opposé à *analogique*), ainsi que des procédés utilisant ce mode de représentation (. . .) ».

Références bibliographiques

Candito, M.-H., Nivre, J., Denis, P. & Henestroza Anguiano, E. (2010). Benchmarking of Statistical Dependency Parsers for French. *Proceedings of The 23rd International Conference on Computational Linguistics (COLING 2010)*. Beijing, China.

Gries, S.T., Hilpert, M. (2008). The identification of stages in diachronic data : variability-based neighbour clustering. Corpora 3 (1), pp. 59-81.

Gries, S.T., Hilpert, M. (2012). Variability-based neighbor clustering : a bottom-up approach to periodization in historical linguistics, in : Nevalainen, T., Traugott, E. (eds.), The Oxford handbook of the history of English, Oxford : Oxford University Press, pp. 134-144.

Heiden, S., Tournier, M. (1998). Lexicométrie textuelle, sens et stratégie discursive, in : Actes I Simposio Internacional de Análisis del Discurso, Madrid.

L*e Petit Robert de la langue française* (2016). Paris : Dictionnaires Le Robert. [version en ligne]

Nivre, J., Hall, J., Nilsson, J. (2006). MaltParser : A Data-Driven Parser-Generator for Dependency Parsing. *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*. Genoa, Italy.

Stefanowitsch, A., Gries, S.T. (2003). Collostructions : Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, 8.2, 209-43.

Session 6.B.
Enseignement Langue Étrangère sur corpus (Oral)

Formulaic language in the EFL classroom: a corpus-based study of phraseological items in British English and American English conversation with implications for EFL teaching

Anna Fankhauser
University of Osnabrück, Institute for English and American Studies
afankhauser@uni-osnabrueck.de

Though long considered a peripheral linguistic phenomenon with little or no relevance for the foreign language learner, phraseology has recently become more of a priority in foreign language teaching research (cf. Martinez/Schmitt 2012: 299). Whereas the term phraseology has traditionally been used to refer exclusively to narrowly defined lexical items such as idioms and proverbs, its definition has been widened to allow for a more general understanding of phraseological items as pragmatically fixed or formulaic units (cf. Stein 2007: 220).

Such relatively fixed multiword expressions offer considerable advantages for the language learner. Research suggests that formulaic units are stored and retrieved as wholes from the brain and that the language user does not need to infer their meaning by analysing each individual component (cf. Ellis 1996, Martinez/Schmitt 2012, Sinclair 1991, Wray 2002 & 2008). This leads to increased fluency in text production as well as to greater efficiency in language use, processing and acquisition (cf. Wray 2002, Nation/Shin 2007).

The significance of formulaic language for foreign language learning is now widely accepted and a considerable amount of research has been done on teaching phraseology in the EFL classroom. Nattinger and DeCarrico, for example, established a classification of language chunks and discussed potential applications for pedagogy (Nattinger/DeCarrico 1992). Other researchers have focused on phraseology for specific purposes: Simpson-Vlach and Ellis, for example, created a list of formulaic sequences from academic contexts (Simpson-Vlach/Ellis 2010). Yet, we still lack teaching materials and a systematic approach to equipping language learners with the necessary general phraseological knowledge. Schmitt and Martinez were the first to systematically address this problem in 2012, when they published their corpus-based "PHRASal Expressions List (PHRASE List), a list of the 505 most frequent non-transparent multiword expressions in English" (Martinez/Schmitt 2012: 299). The PHRASE List was intended as a basis for the systematic incorporation of multiword expressions into language pedagogy. The study limits itself to the first 5,000 word families in the British National Corpus (BNC), thus taking into account all n-grams occurring at least 787 times in the BNC. As it was the authors' ambition to exclude items with little or no semantic content (*e.g. is the*), Schmitt and Martinez established additional criteria to refine the list.

The present study seeks to expand and refine this approach by establishing a comprehensive corpus-based list of high- and mid-frequency phrasemes found in spoken British and American English. It aims to categorize the items by function and pragmatic context. Moreover, it will describe previously unnoticed or insufficiently investigated items using lexicographic criteria. This will provide the basis for the systematic teaching of relevant phraseology to English language learners. The overall objective of the list is to offer a viable alternative approach to teaching phraseological items that goes beyond merely subjective approaches such as teachers' individual preferences or the random occurrence of phraseological items in textbooks.

Although British English is the dominant language variety in the German EFL classroom, the present study will also take account of American English phraseology, as American English is probably equally relevant to learners. This is why the project has been based on two different corpora that have been carefully balanced in order to achieve maximum comparability between them and to avoid any distortion of the results and the conclusions drawn from them. To this end, an approximately equivalent number of items was included in the corpora (around 150 million) and the proportions of the various text types were matched evenly.

Following recent studies (e.g. Quaglio 2009, Levshina 2017) which suggest that film and television language does not fundamentally differ from naturally-occurring conversation, the present research project uses television dialogue and subtitles as its primary data. The corpus used to investigate British English consists of subtitle data from BBC TV programmes, such as documentaries, movies,

news programmes, reality shows, TV series and sports programmes. The American corpus is based on the spoken part of the Corpus of Contemporary American English (COCA), which consists of TV and radio programmes, such as news formats on different topics, discussion shows, documentaries, interviews, talk shows or call-in shows, as well as of film scripts and transcripts of American TV series.

To establish a fairly comprehensive list of spoken phrasemes, the threshold level for the study had to be lower than that set by Schmitt and Martinez. According to Paul Nation (2006), a vocabulary of at least 6,000 to 7,000 word families is needed to adequately understand spoken texts, which roughly corresponds to 28,000 individual words (cf. Schmitt 2008: 232). Consequently, phrasemes occurring as often as, or more often than, the 28,000th individual word in each corpus were considered to be worth teaching to EFL learners. In order to facilitate the list's direct pedagogical use, an additional set of criteria was established for selecting the phrasemes to be incorporated into the list:

- N-grams carrying little or no semantic content (e.g. *of the, in the*) as well as incomplete expressions (e.g. l*ot of*) were eliminated.
- In addition, word combinations reflecting valency structures or colligations (e.g. *you want to*) were excluded from the list.
- Compositional expressions not displaying any pragmatic fixedness (e.g. *the fact*) as well as compound nouns (e.g. *health care*) were also removed from the list.
- The common defining criteria of phrasemes, polylexicality as well as syntactic, lexical and/or pragmatic fixedness, have to be fulfilled.

To avoid omitting important multiword expressions, items consisting of up to 6 elements were investigated (Schmitt and Martinez limited themselves to investigating 2- to 4-grams). The procedure is thus based on Bürgel's and Siepmann's approach, who worked on identifying the most important phraseological items for French.

As a first step, two n-gram lists were generated using the Sketch Engine, an online-based corpus tool which allows users to analyse openly accessible corpora but also individually compiled corpora. Next, all relevant items were extracted from the two n-gram lists using the aforementioned criteria. The resulting phraseme lists were then reordered by dividing the items into three categories: referential, structural and communicative phrasemes. According to Burger (52015: 31f.), referential phrasemes refer to real or fictional objects, events or circumstances. Structural phrasemes are used to establish syntactic relations, whereas communicative phrasemes allow the speaker to establish communicative relations with the persons involved in a communicative exchange. In a third step, subcategories were developed in order to further define the items' functional characteristics. Communicative phrasemes, for example, were classified into different speech acts such as greeting, agreeing or prompting.

Close analysis of the American English phrasemes found has already yielded valuable insights into multiword expressions that have so far received inadequate treatment in learner's dictionaries, although they should be part of at least the receptive phraseological core vocabulary of EFL learners.

The phraseme *you tell me*, for example, is not mentioned in LDOCE, OALD and CALD, three of the major learner's dictionaries, although 4,018 occurrences in the 144,402,542-word corpus suggest that it is used rather frequently by speakers of American English. Some online dictionaries, however, list the item as a phraseme (see below). In its phraseological meaning *you tell me* is mainly used as a response to questions to which the answer is unknown, e.g.: "What's Shelby got to do with this?" – "You tell me." / "Why would she do that?" – "You tell me." The person using *you tell me* as a response usually wants to emphasize that it is obvious that the question cannot be answered and/or that it can be considered as restated. *Dict.cc* offers two German translations that confirm the conclusions drawn from the corpus query: "Sag du es mir!" and "Das weiß ich genau so wenig wie du!". *Urban Dictionary* provides two meanings for the item. The first definition, which is rated as "top definition", differs from the corpus query results. It is based on a Facebook post commenting on a Facebook status and suggests that you tell me is an "expression of strong agreement". The second meaning given by Urban Dictionary corresponds to the corpus query results. Unfortunately, the user can only draw on an example without a clearly identifiable source instead of a definition: "Pottsy says: what

happened last weekend? - Ed says: You tell me". The third online dictionary that lists *you tell me* is *YourDictionary*. Here, the item is categorized as an interjection: "(in response to a question) I don't know!"

Another phraseme that has gone unrecorded in the dictionaries mentioned above is *you do that*. The item occurs 3,946 times in the corpus for American English and is used in affirmative response to an interlocutor who has just announced plans for the near future, e.g.: "I'll call you back." – "Yeah, you do that." / "I gotta go to the bathroom." – "Good. You do that." / "I guess I'd better wait outside." – "You do that." The item allows the speaker to perform the speech act of prompting and to express encouragement or even command.

I can tell (3,961 occurrences) is another item that is absent from the dictionaries mentioned above, at least in this particular form. It allows the speaker to express subjective certainty, e.g.: "She's odious to him. I can tell. Just like I can tell he likes me." / "I didn't know he could do that. He's been listening to me. I can tell." / "You don't have children, I can tell." / "You were so much in love." – "Joe. I know. I can tell."

It is hoped that this brief introduction shows that the present study will not only provide EFL learners with a fairly comprehensive list of useful phraseological items but that detailed phraseological analysis can assist language learners in making correct use of multiword expressions. The aforementioned categorization and investigation of the phrasemes found in the two corpora will make it possible to give a detailed listing of the pragmatic and lexico-grammatical contexts in which phrasemes are typically embedded, context-sensitive translations as well as colligational preferences. This will ensure that the findings of the study can be applied to individual learning situations by everybody involved in the learning process.

Abréviations

BNC – British National Corpus
CALD – Cambridge Advanced Learner's Dictionary
COCA – Corpus of Contemporary American English
LDOCE – Longman Dictionary of Contemporary English
OALD – Oxford Advanced Learner's Dictionary

References

Online dictionaries

Cambridge Advanced Learner's Dictionary.
Dict.cc (http://www.dict.cc/?s=you+tell+me,19.01.2017).
Longman Dictionary of Contemporary English.
Oxford Advanced Learner's Dictionary.
Urban Dictionary (http://www.urbandictionary.com/define.php?term=you%20tell%20me,19.01.2017).
YourDictionary (http://www.yourdictionary.com/you-tell-me, 19.01.2017).

Other literature

Bürgel, C./Siepmann D. (i. V.). Les unités phraséologiques fondamentales du français contemporain. In: Kauffer, M./Keromnes, Y. (eds.). *Approches théoriques et empiriques en phraséologie.* Tübingen: Stauffenberg.

Burger, H. (52015). *Phraseologie. Eine Einführung am Beispiel des Deutschen.* Berlin: Erich Schmidt Verlag.

Ellis, N. (1996). Sequencing in SLA. Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition*, 18, 91-126.

Levshina, N. (2017). Subtitles as a Corpus. An n-gram approach. *Corpora*, to be pusblished.

Martinez, R./Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics*, 33/3, 299–320.

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review/La Revue canadienne des langues vivantes*, 63/1, 59-82.

Nation, P./Shin, D. (2007). Beyond single words: the most frequent collocations in spoken English. *ELT Journal*, 62/4, 339-348.

Nattinger, J. R./DeCarrico, J. S. (1992). *Lexical Phrases and Language Teaching.* Oxford: Oxford University Press.

Quaglio, P. (2009). Television Dialogue. T*he sitcom Friends vs. natural conversation.* Amsterdam; Philadelphia: John Benjamins.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research*, 12/3, 329-363.

Simpson-Vlach, R./Ellis, N. C. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, 31/4, 487-512.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford: Oxford University Press. Stein, S. (2007). Mündlichkeit und Schriftlichkeit aus phraseologischer Perspektive. In: Burger, H. (ed.). Phraseologie. Ein internationales Handbuch der zeitgenössischen Forschung Vol. I. Berlin; New York: De Gruyter, 220-236.

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge: Cambridge University Press.

Wray, A. (2008). F*ormulaic language: Pushing the boundaries.* Oxford: Oxford University Press.

Du corpus oral à la ressource didactique

Émilie Jouin-Chardon , Carole Etienne et Traverso Véronique
Laboratoire ICAR, ENS de Lyon
prénom.nom@ens-lyon.fr

1. Présentation

Si de nombreux corpus oraux de linguistique ont été collectés, décrits et transcrits pour être mis à disposition de la communauté scientifique à des fins d'enseignement et de recherche, nous constatons que leur utilisation concerne majoritairement le monde de la recherche et peine à se faire une place dans les enseignements de langues. En effet, si les enseignants reconnaissent la richesse des corpus authentiques, ils exploitent encore peu ces réservoirs de données "réellement attestées" que constituent les corpus oraux, leur préférant des ressources audio ou vidéo empruntées au cinéma ou à la télévision.

Les corpus d'interactions variées enregistrées en situation réelle issus de la base de données CLAPI (Corpus de LAngue Parlée en Interaction, http://clapi.ish-lyon.cnrs.fr/, qui compte aujourd'hui 370 enregistrements de situations naturelles d'interaction, regroupés en 60 corpus) peuvent permettre aux apprenants de s'immerger dans le français tel qu'il est effectivement parlé en contexte. L'utilisation de ces corpus permettrait ainsi une ouverture sur le « réel » des activités sociales et des usages du français. Comme l'affirme Debaisieux (2008) : « l'intérêt des données orales authentiques est de permettre la constitution de descriptions pédagogiques plus proches des comportements langagiers effectifs des natifs », ce qui permet en quelque sorte une confrontation de la langue d'usage avec les formes plus standard de la langue qui sont tout naturellement enseignées en classe.

Au-delà des corpus eux-mêmes, c'est aussi l'ensemble des recherches en linguistique de l'oral qui ne bénéficient que trop peu à l'enseignement du français, qu'il s'agisse de documenter des spécificités de notre langue orale ou d'en analyser les récurrences.

Dans le paysage des bases de données de corpus oraux sur le Français, la réflexion sur la mise à disposition de corpus et de résultats de travaux de recherche est une réalité qui concerne la plupart des grandes bases de données en linguistique de l'oral. Le projet PFC (Phonologie du Français Contemporain, http://cblle.tufs.ac.jp/ipfc/), notamment, a déjà relevé ce défi en décrivant la variation du français oral au travers de fiches explicatives illustrées (Racine *et al.* 2012). Les concepteurs de ces bases de données provenant de différentes disciplines de l'oral, mènent actuellement une réflexion commune autour cette question.

2. Linguistique de corpus, linguistique interactionnelle et didactique des langues

Le lien entre la linguistique de corpus et la didactique des langues s'est construit ces vingt dernières années notamment avec l'approche sur corpus (Debaisieux 2009 ; Boulton, Tyne, 2014). Elle a été initiée par Johns (1991) avec le principe du data-driven learning, où l'apprentissage de la langue par les apprenants passe par la manipulation des données orales dans une démarche pédagogique déductive. Depuis lors, plusieurs travaux et expérimentations ont été menés pour étudier la pertinence de l'utilisation des corpus oraux en tant que documents authentiques dans le cadre de l'enseignement des langues. Le projet exploratoire CLAPI-FPIE a permis d'établir que des apprenants de différents niveaux de langue étaient à même de comprendre des séquences issues de ces corpus (Thomas et al 2016). Cette même initiative a également montré que les corpus authentiques permettaient de confronter l'apprenant à la difficulté des co-constructions, c'est-à-dire la manière dont différents locuteurs articulent et agrègent leurs productions. La dimension sociale et culturelle des interactions a également été mise en avant lors de l'exploitation de ces corpus auprès d'apprenants, en terme de découverte du paysage sonore et visuel (Lhote, 1995) des pratiques sociales par un public qui connait peu ou mal les usages et les pratiques sociales françaises.

Les recherches en linguistique interactionnelle, initiées par les travaux fondateurs de Sacks, Schegloff et Jefferson (1974) sur l'organisation de la parole, ont largement rendu compte des pratiques sociales situées. C'est justement dans le but d'analyser les récurrences et les variations situationnelles du Français parlé en interaction que les chercheurs interactionnistes du laboratoire ICAR ont recueilli depuis de nombreuses années la parole spontanée dans des situations variées du quotidien (réunions

de travail, repas de famille, transactions commerciales, visites guidées, démarchages téléphoniques, sessions de jeu vidéo...). La production langagière n'ayant de sens que dans le contexte et la finalité de sa mise en œuvre, c'est donc une conception de la langue comme dynamique et adaptative : « le savoir langagier ne constitue pas un inventaire statique, clos, transférable tel quel d'un contexte à l'autre, mais un système dont les ressources sont mobilisées et se configurent de manière adaptative, flexible en fonction de contingences locales de l'action langagière » (Pekarek-Doehler, 2006, 11).

Notre projet articule ainsi les recherches menées au sein du laboratoire ICAR (équipe Interaction) sur les spécificités du français oral en interaction avec les recherches effectuées en didactique des langues pour développer les compétences langagières des apprenants (Ravazzolo *et al.* 2015). Pour ce faire, notre équipe a peu à peu constitué un réseau de partenaires chercheurs en didactique et enseignants de Français Langue Etrangère, en France et dans les départements de français à l'étranger, afin de collaborer à la conception de ressources didactiques issues de la base CLAPI et de les valider en termes d'adéquation aux besoins des enseignants, de pertinence ou tout simplement de compréhensibilité.

3. Quelles ressources didactiques

Notre projet vise à confronter les apprenants aux formes et usages de la langue effectivement produite par le biais d'extraits d'interactions naturelles, et en documentant leur fonctionnement dans un format directement utilisable par les enseignants.

En pratique, il s'agit d'explorer les interactions sociales de CLAPI pour sélectionner des extraits illustrant des activités et des actions langagières issus de contextes variés (privé, professionnel, commercial, médical, au téléphone, en visioconférence,...) qui peuvent être pertinents pour l'apprentissage de pratiques sociales langagières. Pour rendre compte de la dimension culturelle, on s'intéresse par exemple à la manière dont on peut accueillir quelqu'un à son domicile, refuser une invitation, commander un produit ou donner des consignes en situation professionnelle.

En collaboration avec des étudiants en Master 2 Didactique Des Langues (DDL) à Paris Descartes coordonnés par Florence Mourlhon-Dallies, nous avons construit une plateforme en ligne (http://clapi.ish-lyon.cnrs.fr/FLE) regroupant actuellement une quarantaine de courts extraits vidéo ou audio d'interactions naturelles transcrits, contextualisés et documentés. On présente ainsi pour chaque extrait : les thèmes de conversation principaux et secondaires, les objectifs pédagogiques, le type de discours (formel, informel), les actions langagières, le vocabulaire utilisé, les difficultés lexicales, les autres difficultés possibles, ... ainsi qu'un volet exploitation proposant quelques types d'activités que l'enseignant pourrait utiliser pour construire son cours.

En parallèle de ces extraits, nous cherchons à construire de nouvelles ressources avec nos partenaires didacticiens, enseignants et linguistes appelées "collections" regroupant et explicitant des attestations de phénomènes langagiers spécifiques de l'oral relevant de la syntaxe, de la grammaire ou du lexique comme l'imparfait de politesse, le futur simple, des expressions comme "c'est vrai que" "tu sais" "tu vois" "en même temps", l'usage de "truc" ("le truc c'est que", "c'est ça le truc", "tu vois le truc", ...), ...

Au-delà de la constitution de ressources didactiques à partir des corpus, notre équipe souhaite rendre également accessibles aux enseignants, sous une forme expliquée et illustrée, les résultats de leur recherche sur l'organisation de l'interaction, sur la variation situationnelle et culturelle ou sur les formes et particularités du français parlé telles que les co-constructions, les dislocations, les répétitions, les reformulations, l'usage de certaines particules ou les disfluences.

4. Perspectives

Cette communication problématisera ainsi l'ajout d'un volet didactique à une banque de données de corpus oraux utilisée principalement par des linguistes, et proposera des solutions pour transposer les résultats des recherches en linguistique interactionnelle en ressources répondant aux besoins des enseignants de français et de linguistique française.

Références bibliographiques

Boulton, A. (éd.) (2009). Des documents authentiques oraux aux corpus : questions d'apprentissage en didactique des langues, *Mélanges CRAPEL*, 31.

Boulton, A. & Tyne, H. (2014). *Des documents authentiques aux corpus. Démarches pour l'apprentissage des langues.* Paris : Didier.

Debaisieux, J-M. (2008). Corpus oraux et didactique des langues : un rendez-vous à ne pas manquer. *Le français dans le monde : recherches et applications. Quel oral enseigner, cinquante ans après le Français fondamental ?* Paris : CLE international FIPF, 102-114.

Debaisieux, J-M. (2009). Des documents authentiques oraux aux corpus : un défi pour la didactique du FLE. *Mélanges CRAPEL*, n°31, 35-56.

Garfinkel, H. (1984). *Studies in Ethnomethodology*, Polity Press : Oxford

Lhote, E. (1995). *Enseigner l'oral en interaction.* Paris : Hachette.

Pekarek Doehler, S. (éd.), (2006). Compétences et langage en action. *Bulletin VALS/ASLA*, 84 (La notion de compétence : études critiques), 9-45.

Racine, I., Detey, S., Zay, F., Kawaguchi, Y. (2012). Des atouts d'un corpus multitâches pour l'étude de la phonologie en L2 : l'exemple du projet « Interphonologie du français contemporain » (IPFC). In : Kamber, A., Skupiens, C. (éds). *Recherches récentes en FLE.* Berne : Peter Lang, 1-19

Ravazzolo, E., Traverso, V., Jouin, E., Vigner, G. (2015). *Interactions, dialogues, conversations : l'oral en français langue étrangère.* Paris : Hachette FLE.

Sacks, H., Schegloff, E. A., Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation, *Language*, 50, 696-735

Thomas, A., Granfeldt, J., Jouin-Chardon, E., Etienne, C. (2016). Conversations authentiques et CECR : compréhension globale d'interactions naturelles par des apprenants de FLE, *Cahiers de l'AFLS* 20(2), 1-44

Comment familiariser les apprenants de français langue étrangère aux fonctionnements de l'oral ? Retour sur la construction d'un corpus oral à visée pédagogique

Christian Surcouf et Alain Ausoni
École de Français Langue Étrangère, Faculté des Lettres, Université de Lausanne, Suisse
prenom.nom@unil.ch

1. Les défis de la description de l'oral en FLM et en FLE...

Dans son article *Qu'est-ce qu'un bon exemple (oral) ?*, Cappeau (2010 : 120) rapporte que « la constitution de corpus oraux a trouvé, en France, une certaine légitimité ». Il se demande toutefois si « cet affichage de l'oral a [...] dépassé le cercle restreint des techniciens de la langue [et s'il a] quelque peu modifié les perspectives adoptées dans des ouvrages à destination du grand public ». L'observation d'exemples de *La grammaire rénovée du français* (Wilmet 2007) et de *Le bon usage* (Grevisse & Goosse 2008) le conduit à conclure que « [l]es exemples oraux [...] continuent de dérouter ou de poser problème, [...] montr[ant] que l'accès à des corpus oraux n'est pas encore aussi simple ni aussi aisé que les habitudes de travail d'une petite communauté linguistique le laissent penser » (Cappeau 2010 : 124). Il n'est pas sûr qu'il s'agisse seulement d'un problème d'accès aux corpus oraux [1], la tradition grammaticale française tendant en effet à considérer que :

> lorsqu'il est question de la langue française, de sa grammaire et de son lexique, c'est en général de la langue écrite qu'il s'agit. La grammaire et le lexique de langue parlée n'apparaissent dans les ouvrages de référence, la plupart du temps, que comme des curiosités marginales dignes d'un petit musée des horreurs de la langue. (Blanche-Benveniste 2003 : 317)

Quand ils ne sont pas entièrement ignorés des grammaires, certains usages oraux s'écartant de la norme écrite se voient plus ou moins implicitement discrédités au travers de libellés tels que *familier* ou *populaire*. Les grammaires rédigées par des linguistes ne sont pas épargnées.

À titre d'illustration, prenons l'exemple de l'usage de *on*, qualifié de « familier » dans *La grammaire méthodique du français* : « La première personne du pluriel (*nous partons*) est fréquemment remplacée à l'oral, familier surtout, par *on* (*on part*) » [2] (Riegel *et al.* 1994 : 34). La priorité accordée à l'écrit transparait dans l'ordre même de la formulation [3] : *nous*, forme attendue (celle de l'écrit), serait « *remplacée* à l'oral » par *on*. Pourtant, si avec Blanche-Benveniste (2003 : 317), on reconnait que « c'est [...] sous sa forme parlée que la langue est le plus largement partagée » et par ailleurs « qu'on apprend à parler avant d'apprendre à écrire » (Saussure (de) 1994 : 47), alors la prise en compte des pratiques de l'oral s'avère fondamentale, et ce quel que soit le prestige accordé à l'écrit dans nos sociétés. En ce qui concerne l'usage du *on*, sur la base de leur « corpus of Everyday Conversational European French (ECEF) » (194 000 mots), Waugh & Fonseca-Greber (2002 : 117) établissent que « 99% of the uses of 1st Pl. tokens [n=1348] are *on*, not *nous* », conduisant les auteures à conclure que « *Nous*, the 1st Pl. of written French, is no longer applicable to a discussion of spoken French. In its stead, the form *on* has undergone a change in its basic meaning, which is now the personal 'we' » (Waugh & Fonseca-Greber 2002 : 125). Dans les descriptions du français des ouvrages de référence, ce résultat statistique devrait logiquement conduire à une inversion de la formulation de Riegel *et al.* (1994 : 34) : « à l'écrit, le *on* est souvent remplacé par *nous* ».

De telles divergences s'avèrent courantes entre les résultats des recherches sur corpus oraux et les descriptions fournies par les ouvrages de référence. On peut par ailleurs regretter qu'à ce jour

1. L'oral étant omniprésent dans le quotidien des auteurs, il est toujours possible de procéder à des relevés ponctuels comme l'ont fait par exemple Damourette & Pichon (1911-1927).

2. L'édition de 2009 modifie seulement l'exemple : « on va au ciné ? » (Riegel *et al.* 2009 : 62).

3. Tension que résume ainsi Linell (2005 : 29) : « we can talk about a paradox in modern linguistics : one claims the absolute primacy of spoken language, yet one goes on building theories and methods on ideas and experiences of a regimented, partly made-up language designed for literate purposes and overlaid with norms proposed by language cultivators, standardisers and pedagogues. All this amounts to a deeply ingrained contradiction based on a veritable reversal of priorities ».

aucun équivalent français d'une *Grammar of Spoken and Written English* n'existe, qui adopterait « a corpus-based approach, which means that the grammatical descriptions are based on the patterns of structure and use found in a large collection of spoken and written texts, stored electronically, and searchable by computer » (Biber *et al.* 1999 : 4). Un tel ouvrage constituerait non seulement une ressource scientifique précieuse pour la linguistique, mais présenterait de surcroit l'avantage d'établir une base de référence pour tout grammairien ou concepteur de manuels pédagogiques de français, et plus particulièrement de français langue étrangère (FLE). En effet, si le locuteur francophone natif maitrise par définition les fonctionnements oraux qu'il pratique dans son quotidien depuis son enfance, tel n'est pas le cas pour l'apprenant allophone. Sa compréhension orale passe nécessairement par l'appropriation des fonctionnements caractéristiques du français parlé. Or, les grammaires et les manuels de FLE s'inscrivent eux aussi dans la continuité des descriptions traditionnelles, et ne font qu'une place marginale à l'oral. Ainsi, pour l'usage du *on*, évoqué plus haut, la *Nouvelle Grammaire du Français*, destinée à un public de FLE, stipule que « dans la langue familière, le pronom indéfini *on* s'emploie comme un pronom personnel à la place de *nous* » (Delatour *et al.* 2004 : 73). Dans la préface de leur ouvrage, les auteurs évoquaient pourtant leur « ambition de fournir aux apprenants de français langue étrangère un manuel qui leur donne des repères précis pour maitriser l'expression écrite et orale » (Delatour *et al.* 2004 : 3). Comment l'apprenant de FLE parviendrait-il à « maitriser [...] l'expression orale » et la compréhension orale, si les descriptions des auteurs ne reflètent pas les pratiques effectives mises en évidence par la linguistique de corpus [4] ? Il ne s'agit pas là d'un exemple unique. Bien d'autres caractéristiques courantes de l'oral – l'usage de la négation sans *ne* [5], la réduction de *tu* à /t/ devant voyelle [6], la chute du /l/ de *il* devant consonne [7], les dislocations, etc. – sont soit ignorées, soit disqualifiées, ou condamnées par ces ouvrages. On aurait pu espérer que les enregistrements des manuels de FLE compenseraient de telles lacunes en offrant une représentation fidèle du français parlé, mais tel n'est pas le cas. Les documents sonores des manuels sont le plus souvent construits et enregistrés en studio à partir d'une base écrite, les rendant dès lors peu représentatifs de l'oral tel qu'il est pratiqué au quotidien par des millions de francophones natifs (voir à cet égard les recherches de Bento 2007 ; Giroud & Surcouf 2016 ; Surcouf & Giroud 2016 ; Vialleton & Lewis 2014).

En somme, l'apprenant de FLE se heurte à deux obstacles dans sa compréhension du fonctionnement de l'oral. Inscrits dans une longue tradition grammaticale valorisant l'écrit, les manuels ou les grammaires de FLE ne fournissent guère d'informations pertinentes sur le français parlé. Par ailleurs, si elle existe, l'information, dispensée sous sa forme écrite – donc silencieuse –, occulte la dimension sonore constitutive de l'oral [8], conduisant parfois à l'usage d'artifices graphiques à l'instar de la flèche ci-dessous, censée renvoyer à un contour prosodique précis, que l'apprenant devrait idéalement parvenir à reproduire, même en l'absence d'exemple sonore.

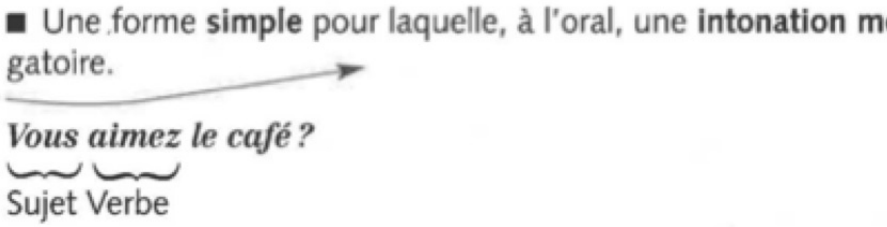


FIGURE 1 – (Poisson-Quinton et al. 2003 : 95)

Dès lors, comment sensibiliser les apprenants de FLE aux caractéristiques du français parlé en

4. À elle seule, dans l'exemple de l'usage du *on*, l'intuition d'un francophone, si elle n'était pas conditionnée par une forte tradition grammaticale fondée sur des normes écrites, devrait permettre de parvenir à la même conclusion.

5. Par exemple, Glaud *et al.* (2015 : 163) écrivent : « Les éléments *ne* et *pas*, *plus*, *rien*, *personne* ou *jamais*, permettent d'exprimer une négation ». Aucune mention n'est faite de l'usage sans *ne* de l'oral.

6. Ce phénomène est signalé et condamné dans Delatour *et al.* (1991 : 163) : « ne dites pas "T'aimes", dites "Tu aimes" ».

7. Dans leur *Grammaire expliquée du français*, Poisson-Quinton *et al.* (2003 : 34) écartent implicitement une telle prononciation : « *Ils*, *elles*, se prononcent [il] et [ɛl] devant une consonne : *Ils parlent* [ilparl] ; *elles parlent* [ɛlparl] », et ce contre les résultats des études sur corpus (voir Ashby 1984 ; Brognaux & Avanzi 2015).

8. « La notation de l'oral par écrit constitue une contradiction irréductible : l'écrit ne présentera jamais qu'une image approximative de la réalité linguistique orale » (Gadet 1989 : 43).

évitant les écueils exposés ci-dessus ?

2. La construction d'une base de données orales à visée pédagogique

En guise de réponse, nous avons entrepris de constituer une base de données orales (en cours d'élaboration) à partir de reportages et d'entretiens diffusés sur France Culture, et renvoyant à des situations de communication variées (par ex. *Les pieds sur terre*, *À voix nue*, *Tout un monde*, etc.). Alors que les reportages – mixés et formatés pour une diffusion en différé – offrent un panorama étendu du français quotidien parlé par toutes les catégories socioprofessionnelles, les entretiens et débats présentent quant à eux un oral proche de celui entendu par nos étudiants dans le cadre universitaire.

Chaque émission – sélectionnée pour son intérêt pédagogique – est transcrite, alignée, et annotée à l'aide du logiciel Elan [9]. À ce jour, l'annotation couvre une centaine de phénomènes phonétiques, syntaxiques, lexicaux et conversationnels. Chacun des phénomènes sera cherchable via une interface reprenant ces quatre dimensions. Pour un phénomène donné, l'apprenant pourra accéder à toutes les occurrences présentes dans le corpus, et les écouter à sa guise, tout en s'étayant sur la transcription et les informations de la notice pédagogique ciblant les niveaux A2 à C2. À titre d'illustration, présentons quelques-uns des phénomènes phonétiques apparaissant dans l'énoncé suivant, extrait d'un reportage diffusé dans *Les pieds sur Terre* sur France Culture :

> **(1) il devait** y avoir **(2) je sais pas** dix personnes un truc comme ça les amis proches
> et ceux que (**3) ça faisait** rigoler **(4)** d'**être là** (FC_PST_2016.04.27_193-194)

Ici, l'apprenant pourra prendre connaissance de chacun des phénomènes signalés en gras, soit en (1) la réduction de /il/ à /i/ [idve], en (2) l'assimilation de /ʒ/ à /ʃ/ et la réduction [ʃepa] qui s'ensuit, en (3) la chute du /ə/ conduisant à la prononciation [savze], et enfin en (4) la chute du /ʁ/ final de /ɛtʁ/ [dɛtla]. Pour chacune de ces annotations, l'apprenant aura accès à la transcription, à l'extrait sonore, et aux explications pédagogiques du phénomène annoté et surligné en couleur dans la transcription en orthographe conventionnelle fournie dans l'interface.

Les fonctionnalités de cette interface se démarqueront ainsi de celles disponibles sur des sites comme *Sacodeyl* [10] (université de Murcia) ou *Backbone* [11] (université de Tübingen), qui, bien qu'ils poursuivent également des objectifs pédagogiques sur la base de corpus oraux, offrent des « possibilités de recherche [qui] sont en fait les mêmes que pour les corpus écrits (avec éventuellement la possibilité d'écouter le son pour des segments donnés du corpus) » (Boulton & Tyne 2014 : 50).

Dans notre communication, nous évoquerons dans un premier temps les raisons pour lesquelles nous avons entrepris de construire ce corpus de français parlé à visée pédagogique, interrogeable par des apprenants de FLE (A2 à C2). Nous présenterons ensuite l'interface informatique et son usage à la fois du point de vue du concepteur, de l'administrateur, et de celui de l'usager-apprenant.

Références bibliographiques

Ashby, W. J. (1984). The Elision of /l/ in French Clitic Pronouns and Articles. *Romanitas. Studies in Romance Linguistics*, 4, 1-16.

Bento, M. (2007). Le français parlé : une analyse de méthodes de français langue étrangère. In : Abecassis, M., *et al.* (Eds.) : Le français parlé au 21ème siècle : Normes et variations dans les discours et en interaction. Annales du Colloque d'Oxford (juin 2005). Volume 2. Paris : L'Harmattan, 191-212.

Biber, D. ; Johannson, S. ; Leech, G. ; Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow : Longman.

Blanche-Benveniste, C. (2003). La langue parlée. In : Yaguello, M. (Ed.) : Le Grand Livre de la Langue française. Paris : Seuil, 317-344.

Boulton, A. & Tyne, H. (2014). *Des documents authentiques aux corpus. Démarches pour l'apprentissage des langues*. Paris : Didier. item[] Brognaux, S. & Avanzi, M. (2015). Sociophonetics of phonotactic phenomena in French. *Proceedings of the International Conference on Phonetic Sciences* (ICPhS), 1-5.

9. Ce projet ayant débuté au cours de 2016, une dizaine d'heures ont été transcrites et alignées. L'annotation des phénomènes est en cours.

10. http ://sacodeyl.inf.um.es/sacodeyl-search2/

11. http ://webapps.ael.uni-tuebingen.de/backbone-search/faces/search.jsp

Cappeau, P. (2010). Qu'est-ce qu'un bon exemple (oral) ? *Travaux linguistiques du Cerlico*, 23, 119-132.

Damourette, J. & Pichon, É. (1911-1927). *Des Mots à la Pensée, Tome I*. Paris : D'Artrey.

Delatour, Y. ; Jennepin, D. ; Léon-Dufour, M. & Teyssier, B. (2004). *Nouvelle Grammaire du Français*. Paris : Hachette FLE.

Delatour, Y. ; Jennepin, D. ; Léon-Dufour, M. ; Mattlé-Yeganeh, A. & Teyssier, B. (1991). *Grammaire du français*. Paris : Hachette FLE.

Gadet, F. (1989). *Le français ordinaire*. Paris : Armand Colin.

Giroud, A. & Surcouf, C. (2016). De « Pierre, combien de membres avez-vous ? » à « Nous nous appelons Marc et Christian » : réflexions autour de l'authenticité dans les documents oraux des manuels de FLE pour débutants. *Congrès Mondial de Linguistique Française*, 2016, 1-18.

Glaud, L. ; Lannier, M. ; Loiseau, Y. ; Perrard, M. & Rimbert, O. (2015). *Grammaire essentielle du francais A1-A2*. Paris : Didier.

Grevisse, M. & Goosse, A. (2008). *Le Bon Usage. Grammaire française*. Bruxelles : De Boeck/Duculot.

Linell, P. (2005). *The written language bias in linguistics : its nature, origins, and transformations*. New York : Routledge.

Poisson-Quinton, S. ; Huet-Ogle, C. ; Boulet, R. & Vergne-Sirieys, A.-. (2003). *Grammaire expliquée du français. Niveau débutant*. Paris : Cle International.

Riegel, M. ; Pellat, J.-C. & Rioul, R. (1994). *Grammaire méthodique du français*. Paris : Presses Universitaires de France.

Riegel, M. ; Pellat, J.-C. & Rioul, R. (2009). *Grammaire méthodique du français*. Paris : Presses Universitaires de France.

Saussure (de), F. (1916/1994). *Cours de linguistique générale*. Paris : Payot.

Surcouf, C. & Giroud, A. (2016). À quelle langue accède l'apprenant ? Examen critique du traitement de l'oral dans les premières leçons de manuels de français langue étrangère pour débutants. *Linguistik Online*, 78-4, 11-27.

Vialleton, É. & Lewis, T. (2014). Reconsidering the authenticity of speech in French language teaching : theory, data, methodology, and practice. In : Tyne, H., *et al.* (Eds.) : French through Corpora : Ecological and Data-Driven Perspectives in French Language Studies Newcastle upon Tyne : Cambridge Scholars Publishing, 293-316.

Waugh, L. R. & Fonseca-Greber, B. (2002). Authentic materials for everyday spoken French : corpus linguistics vs. French textbooks. *Arizona Working Papers in SLAT*, 9, 114-127.

Wilmet, M. (2007). *Grammaire rénovée du français*. Bruxelles : De Boeck.

Logiciel : Elan : Max Planck Institute for psycholinguistics Nijmegen : http ://tla.mpi.nl/tools/tla-tools/elan/

Session 7.A.
Analyses lexicologiques

Quels outils pour l'étude de la variation du français ? L'apport de la linguistique de corpus à l'exemple d'un diatopisme polysémique, *prime* (adj.)

Inka Wissner
Old Dominion University
inkawissner@gmx.net

La linguistique française dispose désormais de nombreux corpus préparés au cours des dernières décennies, établis avec des objectifs variables. Ces outils deviennent de plus en plus incontournables pour décrire le français à partir de réalisations effectives de la langue ; en même temps, ils imposent au chercheur de maîtriser les outils de la linguistique de corpus et de prendre un ensemble de précautions (cf. Cappeau/Gadet, 2007). Pour décrire la langue générale, des projets comme le TLF ont ainsi démontré la valeur ajoutée d'un dictionnaire qui s'appuie exclusivement sur un corpus (Frantext), plutôt que sur ses prédécesseurs. Pourtant, en linguistique variationnelle, et en lexicographie francophone en particulier (où *francophone* est utilisé au sens de « qui porte sur le français dans l'espace francophone »), la variation spatiale est encore largement identifiée selon une approche différentielle par rapport à un ensemble métalinguistique. Cette approche consiste pour l'essentiel à exploiter les ouvrages de référence qui font autorité, comme le TLF et FEW et les dictionnaires d'états anciens de la langue (*français de référence* selon Poirier, 2005 : 497). Ceux-ci sont complétés de sources spécialisées qui portent sur des domaines mal ou non représentés en lexicographie générale, en particulier sur des langues en contact et la variation diatopique du français même (comme DSR, DHFQ, DRF, DLF ; BDLP). Ces derniers dictionnaires s'appuient quant à eux sur les données lexicographiques antérieures, des enquêtes de terrain, ou encore des ensembles textuels, écrits et oraux. Cet ensemble reste métalinguistique, mais constitue un outil de travail objectivable, contrairement à l'idée abstraite que le lexicographe peut se faire du français général, ou même d'un 'bon usage', voire d'un français 'de Paris' ou 'de France' – qui eux aussi s'intègrent dans un diasystème variable (cf. Coseriu, 1981).

Dans l'approche différentielle, les corpus sont donc bien utilisés pour analyser l'une ou l'autre variété de français, mais ont pour l'essentiel un rôle complémentaire, venant diversifier la nature de données majoritairement lexicographiques. L'existence des méthodes de la linguistique de corpus, le nombre croissant de ressources textuelles, leur taille et leur présence dans différentes zones de l'espace francophone ne permettraient-ils pas d'identifier ce qui relève de la variation diatopique dans la langue par une comparaison *pan*francophone des corpus ? Dans une telle optique, quel est l'apport effectif des ressources textuelles ? Permettent-elles de se passer de la documentation lexicographique traditionnelle ?

Pour répondre à cette problématique, cette communication recourt aux concepts et outils de la lexicologie philologique historique, et de l'approche comparative panfrancophone en particulier. Celle-ci est complétée par la méthode fine de l'étymologie-reconstruction, qui vise à reconstruire progressivement les étapes d'évolution d'un élément de la langue. Elle s'appuie sur une comparaison des régularités formelles et les évolutions sémantiques parallèles des formes d'une famille lexicale, et donc sur les traces laissées par la documentation historique, mais aussi sur la documentation contemporaine – qui reste la plus vaste et la mieux établie (Chauveau, 2013 : 178-179).

L'approche lexicographique traditionnelle sera dans cette communication confrontée à une exploitation de données textuelles dans une visée panfrancophone. Je constituerai pour cela un ensemble cohérent de corpus, en comprenant *corpus* comme un ensemble de réalisations discursives qui a été conçu pour des fins d'analyse linguistique dans le respect de critères de sélection particuliers et qui est exploitable par l'intermédiaire d'un moteur de recherche. J'ai sélectionné les plus grands corpus du français contemporain monolingues dans la francophonie qui dépassent une taille d'environ 5 millions de mots occurrences, qui permettent des requêtes lexicales, et qui sont rendus accessibles en ligne à la communauté scientifique. Ce corpus est susceptible de permettre une analyse efficace, donc à la fois fructueuse et rentable, même pour des unités lexicales marquées. S'il est sélectif, il a pour grand avantage de constituer un outil de travail vérifiable par d'autres membres de la communauté scientifique, pouvant servir de corpus 'de référence', comme un pendant textuel des dictionnaires et grammaires de référence.

Si la taille des corpus n'est pas un critère satisfaisant, il est pertinent surtout pour l'analyse de noms, verbes, et adjectifs – donc d'unités qui appartiennent aux classes majeures, comme celles analysées ici – moins fréquents dans l'absolu que les unités appartenant aux classes mineures, comme les conjonctions ou les prépositions. Le seuil retenu ici est donc minimal. Même parmi les spécialistes de corpus qui s'intéressent non pas à des dimensions variationnelles de la langue, mais à la langue générale, la taille actuelle des corpus oraux (autour de quelques milliers de mots), ne permet pas « de faire des recherches lexicales ni d'établir des statistiques fiables sur les usages » (Baude, 2006 : 29). Selon les standards internationaux, un corpus d'étude conforme donnerait accès à 3 millions de mots à l'oral, 15 millions à l'écrit (ORFÉO, en cours). Dans une visée diatopique, le critère de la taille est d'autant plus critique au sens où il doit s'appliquer non pas à l'ensemble des données exploitées, mais à chaque zone de l'espace francophone. Il s'agit donc bien d'un des critères de valeur pour permettre des analyses lexicales qualitativement fructueuses sur la variation diatopique, à côté de paramètres comme la spécificité des sujets et des genres discursifs qu'ils accueillent, et les types de requêtes que permettent leurs moteurs de recherche (Wissner, 2012 : 252). Les limites d'exploitation de corpus oraux pour des recherches lexicales dans une visée diatopique avaient déjà été soulevées dans le cadre de la description des diatopismes du français en Belgique et en Afrique (Queffélec, 1997). La situation a-t-elle considérablement changé en vingt ans ?

Pour répondre à cette question, nous exploitons ici des corpus qui sont de trois types. Il s'agit tout d'abord des corpus traditionnels qui donnent accès à des données surtout littéraires : Frantext (francophonie) et FLI (Canada) ; s'y joignent l'*Est Républicain*, pour des données journalistiques de l'est de la France, et Varitext, pour des données journalistiques et littéraires en Europe et en Afrique. Un deuxième type de corpus contient des données tirées du web francophone : FrWac – devenu le plus grand corpus de la langue française – et I-FR ; ils donnent accès à des données variées en termes thématiques, situationnels et discursifs, et apportent des données de types de discours propres (comme les blogs). Un troisième type est celui des corpus de transcriptions d'enregistrements oraux, où seule la base ESLO répond au seuil quantitatif fixé (France : Ouest). Approchant les sept millions de mots selon des approximations, c'est désormais le corpus oral le plus grand du français qui soit entièrement rendu accessible à la communauté scientifique. Le corpus belge VALIBEL reste en effet temporairement indisponible. Le seuil retenu amène aussi à exclure des corpus d'enregistrements pour différentes zones de la francophonie qui sont comparables grâce à l'application d'une méthodologie unifiée (PFC, CIÉL-F). L'essentiel de leurs données n'est pas encore librement accessibles ; ils ne sont donc consultés qu'à titre complémentaire. Pourtant, ces corpus seront primordiaux pour faire évoluer les analyses lexicales à visée panfrancophone, tout comme le Corpus d'Étude pour le Français Contemporain (CEFC) en constitution – qui fournira quant à lui un corpus stable pour l'Europe surtout (aussi Amérique du Nord).

Pour l'Amérique du Nord, en attendant le corpus panaméricain FRAN, les besoins de la recherche m'amènent ici à ajouter des ensembles de moindre envergure : le sous-ensemble en accès libre de la BDTS, interrogeable par l'intermédiaire d'un index (comme FLI), ainsi qu'en complément le corpus 'oral' CFPQ et le corpus MCVF pour les états antérieurs du français. Pour la France, s'est enfin vu ajouter le corpus RÉGION, consultable à Nancy seulement. J'exclurai de l'analyse les ensembles textuels qui n'ont pas été conçus pour l'analyse linguistique, comme les ressources journalistiques Europresse et Eureka ainsi que les ressources web instables que sont Gallica, GRL et Google Web. En effet, leur statut de corpus est largement débattu ; certes très utiles pour l'analyse de formes rares, ces ressources ne permettent pas de recherches linguistiques sophistiquées comme l'exige l'analyse d'unités lexicales complexes (que ce soit formellement ou sémantiquement).

Pour vérifier l'apport effectif des corpus dans une optique historico-comparative panfrancophone, ma communication propose une étude de cas plutôt qu'une approche exhaustive. Elle portera sur une forme polysémique du français moderne, *prime* (adj.) (Réf. *précoce ; vif ; soupe au lait* ...). Ce choix ciblé rend possible une étude détaillée qui permet d'évaluer l'apport des corpus pour l'identification de ses différentes caractéristiques : sémantiques, syntagmatiques, géolinguistiques et historiques. L'analyse retracera non seulement les différents emplois de prime en Europe, mais aussi leur trajet historico-variétal à travers le temps et l'espace, d'une variété à l'autre, y compris outre-mer où le français fut durablement exporté à partir du XVIIe siècle. C'est dans un deuxième temps qu'il s'agira

d'apporter des éléments de réponse à la problématique méthodologique soulevée, en évaluant l'apport des dictionnaires de français, d'un côté, et des grands corpus dans l'espace francophone, de l'autre.

Références bibliographiques

Baude, O. (éd.) (2006). *Corpus oraux, guide des bonnes pratiques*. Paris : Éditions du CNRS/Orléans : PU d'Orléans.

BDTS. *Banque de données textuelles de Sherbrooke* recueillant plus de 52 millions de mots de divers types de textes des années 1960 à 2000, préparée sous la direction d'H. Cajolet-Laganière et P. Martel, Sherbrooke (Sherbrooke) ; sous-ensemble de quelque deux millions d'occurrences tirées de 1054 textes, lui-même tiré d'un sous-ensemble de 16 millions de mots, en accès limité par l'intermédiaire d'un index lexical, http://catfran.flsh.usherbrooke.ca/catifq/bdts/index.htm.

Cappeau, P./Gadet, F. (2007). L'exploitation sociolinguistique des grands corpus. Maître-mot et pierre philosophale. *Revue Française de Linguistique Appliquée* XII/1, 99-110.

CEFC (en cours de constitution). Corpus d'Étude pour le Français Contemporain : corpus de référence archivé par Ortolang constituant un outil de recherche sur le français écrit et oral rassemblant des données secondaires à partir de corpus existants ou créés pour le projet ORFÉO via une plate-forme d'interrogation permettant une sélection par méta-données aux ressources proposées et des recherches par requêtes simples et complexes pour permettre de constituer un corpus d'étude conforme aux standards internationaux (3M de mots à l'oral et 15 M de mots d'écrits), en libre accès, http://www.projet-orfeo.fr/ (consulté 02/05/2017).

CFPQ, 2006-, *Corpus de français parlé au Québec* préparé sous la responsabilité de G. Dostie contenant plus de 45 heures d'enregistrements ; support audiovisuel consultable sur place ; transcriptions alignées en accès libre, http://pages.usherbrooke.ca/cfpq/corpus.php.

Chauveau, J.-P. (2013). Fr. *ébarouir* : étymologie-histoire et étymologie-reconstruction. *Revue de linguistique romane* LXXVII, 167-182.

CIÉL-F. *Corpus International et Écologique de la Langue Française* de français oral en interaction rassemblant des extraits d'environ 200 enregistrements de dix minutes collecté de 2006 à 2012 dans des situations comparables dans quinze zones de l'espace francophone ; diffusion libre sur Internet prévue via les interfaces de [moca] et CLAPI (non effective 20/11/2016) (www.ciel-f.org, copyright 2008-2013).

Coseriu, E. (1981 [1958]). Los conceptos de ‹ dialecto ›, ‹ nivel › y ‹ estilo de lengua › y el sentido propio de la dialectología. *Lingüística española actual* III, 1-32.

DHFQ : Poirier, Cl. (éd.) (1998). *Dictionnaire historique du français québécois. Monographies lexicographiques de québécismes*. Sainte-Foy : PU Laval.

DLF : Valdman, A. *et al.* (2010). *Dictionary of Louisiana French : As Spoken in Cajun, Creole, and American Indian Communities*. Jackson : University Press of Mississippi.

DRF : Rézeau, P. (éd.) (2001). *Dictionnaire des régionalismes de France* (DRF). Bruxelles : De Boeck-Duculot.

DSR : Thibault, A. (1997). *Dictionnaire suisse romand. Particularités lexicales du français contemporain. Une contribution au Trésor des Vocabulaires francophones*. Genève : Zoé.

ESLO, *Enquête Sociolinguistique à Orléans*. Corpus oral constitué de 700 heures d'enregistrements menés à Orléans, dont 300 heures de 1968 à 1971 (ESLO-1, à visée didactique) et 400 heures d'enregistrements comparables dans les modalités de collecte, recueillis de 2008 à 2013 (ESLO-2, à visée variationniste) ; accès libre sur Internet sur demande avec signature d'une convention ; accès direct à un sous-corpus anonymisé (350 heures d'enregistrement environ : ca. 200 000 : ESLO-1, 150 : ESLO-2) en ligne, http://eslo.tge-adonis.fr.

Est Républicain. Corpus constitué d'articles de toutes les éditions intégrales du quotidien régional de l'est de la France de 1999, 2002 et 2003 (l'équivalent de deux années pleines), rassemblant un peu plus de 159 millions de mots occurrences (état 6/11/2013), consulté à l'ATILF-CNRS, https://arcas.atilf.fr/cqpweb/ ; consultable sur CNRTL en format XML-TEI P5 en trois fichiers séparés par année, http://www.cnrtl.fr/corpus/estrepublicain/.

Europresse, Archive de textes journalistiques francophones et anglophones en texte intégral, consultable avec accès payant sur Internet, www.europresse.com.

Eureka, Corpus de presse contemporain couvrant l'actualité internationale, nationale, régionale et locale en douze langues dont le français, permettant des requêtes thématiques par mots-clés, rassemblant 1249 sources référencées (état 21/11/2016) ; accès sous contrat, http://www.eureka.cc/Default.aspx, copyright CEDROM-SNi inc. 2016.

FEW : Wartburg, W. von (1928-2003). *Französisches etymologisches Wörterbuch. Eine Darstellung des galloromanischen Sprachschatzes*. Bonn et al. : Klopp et al., 25 volumes.

FLI, *Fichier lexical informatisé*. Base de données du français québécois du XVIe au XXIe siècles comportant 400 000 fiches, en accès libre, http://www.tlfq.ulaval.ca/fichier/.

FRAN (en cours). *Corpus des français d'Amérique du Nord* en préparation sous la responsabilité de F. Martineau visant à établir un ensemble unifié de documents textuels et oraux pour permettre de brosser l'histoire linguistique et culturelle des francophones du Canada et des États-Unis, http://continent.uottawa.ca/corpus-et-ressources-electroniques/ (consulté 03/05/2017).

Frantext, Base de données textuelles de la littérature française : corpus à dominante littéraire constitué de quelque 248 millions de mots du XVIe au XXIe siècles, http://www.Frantext.fr/.

FrWac, Corpus textuel d'une tranche du Web (domaine « .fr ») d'environ 1,6 milliard de mots, construit dans le cadre du projet WaCky Wide Web (Trente/Bologne) ; accès libre sur demande, |urlhttp ://wacky.sslmit.unibo.it/doku.php ?id=download ; exploité sous une version catégorisée conçue à Nancy en collaboration avec Druide Informatique Inc. (Montréal) dont la diffusion via le site CNRTL est prévue.

Gallica, Bibliothèque numérique de la Bibliothèque nationale de France (Paris) rassemblant plus de deux millions de documents, en accès libre http://gallica.bnf.fr/.

GRL, *Google Recherche de Livres*, copyright 2012, http://books.google.com/.

I-FR, Corpus francophone de Leeds d'environ 200 millions de mots tirés de 50 000 pages du Web francophone en 2006 sans limitation de domaines (Leeds), consultable avec recherches de concordances et de collocations, en accès libre, Corpus.Leeds.ac.uk/internet.html.

MCVF, *Modéliser le changement : les voies du français*. Corpus constitué de textes en grande partie intégraux de l'ancien français au français moderne (XVIIIe s.) sous la direction de F. Martineau (Ottawa) ; consultable par mots et cooccurrences ; en accès libre sur demande en cédérom et en ligne, http://www.voies.uottawa.ca/corpus_pg_fr.html, copyright 2009.

PFC, *Phonologie du Français Contemporain : usages, variétés et structure*. Base de données ouverte préparée dans le cadre d'un projet international sous la direction de J. Durand et al. (MoDyCo), recueillies dans différentes zones de la francophonie, consultable par la communauté scientifique en ligne sur demande http://www.projet-pfc.net, copyright 2004-2016.

Poirier, Cl. (2005). La dynamique du français à travers l'espace francophone à la lumière de la base de données lexicographiques panfrancophone. *Revue de linguistique romane*, 69, 483-516.

Queffélec, Ambroise (1997). Le corpus textuel oral. Constitution, traitement et exploitation lexicographique. In Frey, C./Latin, D. (éds.), *Le Corpus lexicographique : méthodes de constitution et de gestion*. Louvain-la-Neuve : Duculot, 353-368.

RÉGION, Banque de données de régionalismes du français hexagonal réunissant 7500 contextes de 220 ouvrages dus à 156 auteurs de France métropolitaine, annoté par région d'appartenance des écrivains ; corpus conçu à Nancy, consultable sur place, Atilf-CNRS/Université de Lorraine.

TLF : Imbs, P./Quemada, B. (1971-1994). T*résor de la langue française. Dictionnaire de la langue du 19e et 20e siècle (1789-1960)*. Paris : Gallimard.

VALIBEL (1989-2009), Banque dynamique et évolutive en ligne de données textuelles du français oral en Belgique rassemblant un ensemble d'environ 4 millions de mots transcrits et alignés, Louvain-la-Neuve, http://www.uclouvain.be/valibel-corpus.html ; indisponible pour une durée indéterminée.

Varitext (2013). *Corpus des variétés nationales du français* préparé par Diwersy S./Mejri, S./Blumenthal, P., constitué de textes journalistiques et romanesques de l'espace francophone en Europe et en Afrique noire, avec environ 80 millions de mots occurrences de textes journalistiques des années 2000 (dont plus de 5 millions sont lemmatisés et syntaxisés). Accès en ligne, http://syrah.uni-koeln.de/varitext/, Cologne/Villetaneuse (20/11/2013).

Wissner, I. (2012). Les grands corpus du français moderne : des outils pour étudier le lexique diatopiquement marqué ?, *SKY Journal of Linguistics* XXV, 233-272.

en fait, c'est quoi, en fait ?

Philippe Martin
LLF (UMR 7110), UFRL, Université Paris Diderot
philippe.martin@linguist.univ-paris-diderot.fr

1. Introduction

La locution *en fait* est devenue très fréquente à l'oral et est communément définie comme locution adverbiale signifiant "*en réalité, effectivement, réellement*". Comme adverbe, ce connecteur se place soit au début ou en fin de phrase, soit avant ou après un verbe ou un groupe verbal, comme dans : ***en fait*** *ce papier n'apporte rien de neuf ; ce papier n'apporte rien de neuf* ***en fait*** *; ce papier* ***en fait*** *n'apporte rien de neuf ; ce papier n'apporte* ***en fait*** *rien de neuf.* De nombreux travaux ont été consacré à cette locution, essentiellement sur les propriétés de sens (par exemple Danjoux-Fluax, 1980 ; Rossari, 1984 ; d'Hondt, 2014).

Toutefois, il s'agit là d'une propriété observée dans l'écrit, qui ne traduirait pas nécessairement les réalités de la production orale, soit parce que les exemples donnés dans les grammaires traditionnelles sont empruntés aux seuls auteurs considérés comme représentatifs d'un certain niveau de langue, soit parce que des positionnements inattendus ne sont attestés que dans l'oral (alors que des positionnements attendus s'y retrouvent rarement). De par son format limité, cette étude se veut plus comme une illustration concrète d'un outil d'analyse optimisé pour les grands corpus oraux que comme une enquête exhaustive sur l'utilisation de *en fait* dans l'oral spontané.

2. Corpus d'analyse

L'écoute des réalisations de *en fait* montre que c'est d'abord l'intonation associée à cette locution adverbiale qui indique à l'auditeur comment l'associer à un autre élément de l'énoncé, groupe verbal, début ou fin d'énoncé, ou autre. Il y a donc lieu pour le savoir d'analyser des corpus oraux, et d'en recenser les regroupements de *en fait* tels qu'indiqués par l'intonation. Pour cette recherche, on a d'une part utilisé la base de données *Orfeo* (2017) qui regroupe un grand nombre de corpus variés d'enregistrements de parole spontanée (CFPB, CFPP, CLAPI, CORALROM, CRFP, FLEURON, ORALNARRATIVE, HUSIANYCIA, OFROM, TCOF, RHAPSODIE, TUFS, VALIBEL), et d'autre part le logiciel d'analyse de la parole WinPitch (2017), qui comporte plusieurs fonctions spécialisées pour l'analyse de grands corpus dont un concordancier intégré.

Ainsi, WinPitch permet de retrouver en un seul clic de souris le segment sonore correspondant à une entrée du concordancier, tout en en affichant instantanément les données acoustiques (dont la courbe mélodique qui est pertinente ici) et ce en convertissant automatiquement les transcriptions dans Orfeo disponibles dans des formats divers (Transcriber, Praat, WinPitch, etc.). Au regard du nombre d'occurrences (5.973 sur 2.995.185 mots), l'utilisation d'un outil d'analyse ergonomique s'avère indispensable.

3. Relations de dépendance

L'analyse des regroupements effectués par la prosodie, et dans lesquels *en fait* est impliqué, s'appuie sur un modèle basé sur les relations de dépendance entre les groupes accentuels, séquences de syllabes ne contenant qu'une seule syllabe accentuée finale. Ces relations sont indiquées par des mouvements mélodiques spécifiques à l'endroit de ces syllabes accentuées, en position finale dans chaque groupe accentuel (Martin, 2009). Ces mouvements sont désignés respectivement par C0 (contour terminal conclusif, déclaratif ou interrogatif), C1 (contour montant supérieur au seuil de glissando), C2 (contour descendant supérieur au niveau de glissando) et Cn (contour neutralisé, montant ou descendant et inférieur au seuil de glissando). C2 détermine une relation de dépendance « à droite » (i.e. vers la suite) envers un contour C1, qui lui-même indique une dépendance « à droite » envers le contour terminal C0. Le contour neutralisé Cn indique une relation de dépendance avec n'importe quel autre contour C2, C1 ou C0 à droite. Enfin le contour C0n similiare à Cn mais de fonction différente indique une relation de dépendance « à gauche » envers le contour terminal C0 qui précède (Martin, 2015). Ces

relations peuvent être résumées par la formule : Cn -> C2 -> C1 -> C0 <- C0n. Les courbes mélodiques associées à chaque exemple révélé par le concordancier permet de déterminer le regroupement de la locution adverbiale avec les autres unités dans l'énoncé.

4. Distribution selon les corpus

Les statistiques relatives aux occurrences de en fait sont données dans la table suivante, ordonnée par nombre d'occurrences croissant :

Corpus	Nombre de mots	Nombre d'occurrences	Pourcentage d'occurrences
ORALNARRATIVE	157.986	18	0,01
RHAPSODIE	65.987	34	0,05
HUSIANYCIA	202.524	180	0,09
VALIBEL	200.381	195	0,10
CFPP	475.078	669	0,13
CORALROM	149.763	303	0,20
CFPB	60.142	127	0,21
OFROM	206.944	480	0,23
TCOF	317.555	765	0,24
CRFP	268.377	674	0,25
TUFS	702.507	1.890	0,27
FLEURON	21.636	66	0,31
CLAPI	166.305	572	0,34
TOTAL	**2.995.185**	**5,973**	**Moyenne 0,20**

Table 1 – Occurrences de la locution en fait dans les différents corpus regroupés dans Orfeo

On a donc relevé 5.973 occurrences de en fait sur un total de 2.995.185 mots. Une première observation porte sur les variations de pourcentage d'occurrences de la locution adverbiale. ORALNARRATIVE ne présente que 0,01% de cas, ce qui s'explique par le style des locuteurs, conteurs professionnels s'adressant à un large public. À l'opposé, le corpus CLAPI contient 0,34% d'occurrences, les enregistrements provenant de locuteurs jeunes dans des situations de repas familiaux ou amicaux. Le corpus CORALROM, mélange de différents styles de parole spontanée (enregistrés en fin des années 1990) révèle un pourcentage de 0,20%, dans la moyenne des différents corpus.

5. Configurations représentatives

Quelques configurations représentatives ont été extraites de ces analyses (les regroupements prosodiques sont indiqués par des crochets) :

5.1 Adverbe en début de phrase, configuration [***en fait*** C1] [phrase C0] :

[*...sept ans et demi exactement euh comment dire* C0] [***en fait*** C1] [*je suis venue au cirque* Cn *un peu par hasard* C1] *parce que euh c'était vraiment les ...* (CRFP [pri-ami-1]) {0 :10.27-0 :12.76}

[*...que c'était c'était pas mal fait c'était bien reconstitué* C0] [**et en fait** C1] [*c'est euh euh Arthur Masson* C1 *est un écrivain ...*] (VALIBEL [famrm1r]) {11 :26.35-11 :29.71}

5.2 adverbe après la fin de la phrase, contour mélodique plat succédant à un contour terminal conclusif, fonctionnant comme thème [phrase C0] [***en fait*** C0n] :

[*... c'étaient des fermiers* C0] [[***en fait*** C0n] (VALIBEL [fames1r]) {24 :44.92-24 :45.91}

5.3 adverbe après la fin de la phrase, contour mélodique terminal conclusif succédant à un premier contour terminal conclusif, fonctionnant comme complément différé (Bally, 1944) [phrase C0] [***en fait*** C0] :

[... [*ben en fait* C1] [*il était concierge* Cn *à Saint-Louis* C0] [***en fait*** C0] [*il a fait plein de petits métiers* ...] (CFPB [cfpb-1000-5] {340.591 s-342.495 s})

[. . . [*ouais mais c'est c'est assez petit* C0] [**en fait** C0] [mais avec mon zoom je pouvais réussir à n'obtenir que ça sur l'image et ...] (VALIBEL [famvv1r]) {2 :37.22-2 :38.62}

5.4 adverbe après la fin de la phrase, contour mélodique terminal conclusif succédant à un contour montant de contimuation majeure, fonctionnant comme ponctuant [phrase C1] [***en fait*** C0] :

[...*je sais pas si tu non c'est un gîte* C1] [***en fait*** C0] [je crois c'est euh le gîte du Saut du Bouchot ou un truc comme ça peut-être ...] (TCOF [anniversaire] {1 :19.23-1 :21.15})

[... *et les vues ouais ouais ouais ouais ouais ouais il voulait récupérer ça* C1] [***en fait*** C0] (CLAPI [reunion_conception_mosaic_architecture]) {55 :29.09-55 :34.38}

Ces quelques exemples montrent que les positions syntaxiques attendues peuvent être différentiées selon les regroupements prosodiques. On note aussi la rareté des exemples en position pré ou postverbale, dont aucun n'a été repris ici.

6. Conclusion

Une première exploration des configurations de la locution adverbiale *en fait* a été réalisée à l'aide du logiciel WinPitch, dans le but d'illustrer par un exemple concret les possibilités d'étude de grands corpus oraux, portant à la fois sur les propriétés syntaxiques et prosodiques. L'ergonomie de ce logiciel permet de recenser et d'analyser dans un temps de travail raisonable les réalisations très variées des locuteurs, illustrant par là le fonctionnement et la pertinence qui regroupements prosodiques qui ne sont pas toujours correctement décrits à partir des seules transcriptions des mots et des syntagmes.

Références bibliographiques

Bally, Ch. (1944). *Linguistique générale et linguistique française*. Berne : Francke.

Danjou-Flaux, N. (1980). À propos de de fait, en fait, en effet, et effectivement. *Le français moderne*, 48, 110-139.

D'Hondt, U. (2014). Au fait, de fait et en fait : analyse de trois parcours de grammaticalisation. *Revue Romane*, 49(2), 235-263.

Martin, Ph. (2009). *Intonation du français*. Paris : Armand Colin.

Martin, Ph. (2015). The Structure of Spoken Language. Intonation in Romance. Cambridge Cambridge University Press.

Orfeo (2017). *Outils et Recherches sur le Français Écrit et Oral*. http://www.projet-orfeo.fr/

Rossari, C. (1992). De fait, en fait, en réalité : trois marqueurs aux emplois inclusifs. *Verbum*, 3, 139-161.

WinPitch (2017). *Logiciel d'anayse acoustique de la parole*. http://www.winpitch.com

Les adjectifs axiologiques dans les guides touristiques : une expérience d'annotation

Jarukan Jitwongnan et Agnès Tutin
LIDILEM, Université Grenoble Alpes
jarukan.jitwongnan@gmail.com
agnes.tutin@univ-grenoble-alpes.fr

1. Introduction

Les guides touristiques constituent un genre discursif riche en adjectifs, visant à « donner un grand nombre d'informations » sur la destination (Maingueneau, 2016, p. 266), voire de conseiller ou prescrire aux lecteurs les attitudes à adopter ou les visites à effectuer. Il est donc courant de trouver des commentaires sur la vie locale, les sites à visiter, les hébergements, etc., dans lesquels l'auteur du guide inscrit son jugement de valeur en utilisant des adjectifs positifs/négatifs (Ex : *un site remarquable, un hôtel correct mais un peu cher*...). Cependant, du fait de la polysémie des adjectifs, il n'est pas toujours évident de déterminer la valeur axiologique ou la polarité de tel ou tel adjectif. En outre, un adjectif « neutre » peut être positif ou négatif en contexte. Afin de repérer les attitudes de l'énonciateur, la définition indiquée dans les dictionnaires paraît souvent insuffisante et il faut effectuer une analyse fine, en se basant sur le corpus et le contexte linguistique. L'annotation que nous souhaitons mener nous permettra de : (i) repérer les asdjectifs axiologiques, y compris les adjectifs non intrinsèquement axiologiques (ii) repérer les mécanismes linguistiques permettant d'identifier l'axiologisation des adjectifs neutres (iii) observer des cibles de jugement et des attitudes du locuteur vis-à-vis de ces cibles, de façon à effectuer une étude préliminaire sur les stéréotypes sur la Thaïlande dans les guides touristiques francophones.

2. Problématique

Les adjectifs axiologiques sont des adjectifs qui indiquent un jugement de valeur et ont une polarité négative ou positive. Kerbrat-Orecchioni (2009) inscrit ce groupe d'adjectifs (ex. *bon, beau*, etc.) dans la catégorie des adjectifs subjectifs évaluatifs qu'elle oppose aux adjectifs objectifs (ex. *célibataire, jaune*, etc.). Les adjectifs axiologiques véhiculent une trace de la subjectivité du locuteur, au sens de la linguistique énonciative, dans la mesure où ils permettent au locuteur d'énoncer son attitude vis-à-vis d'une cible de jugement.

Certains adjectifs portent déjà intrinsèquement un sens axiologique, par exemple *Plusieurs ruines intéressantes* (positif[1]) ; *Un gigantesque bouddha doré très kitsch* (négatif[2]), valeurs que l'on retrouve d'ailleurs précisées dans les dictionnaires. Toutefois, un certain nombre d'adjectifs ne sont pas intrinsèquement marqués mais prennent un sens axiologique en contexte. Par exemple, l'adjectif classique est en général « neutre » au sens propre mais dans, *carte simple de cuisine thaïe classique, mais très bonne* l'adjectif *classique* apparaît plutôt négatif tandis que celui-ci prend une connotation positive dans *Hôtel classique et agréable, en plein coeur de Chinatown.* La définition « neutre » indiquée dans les dictionnaires apparaît donc insuffisante pour vérifier si un adjectif est axiologique en contexte.

L'analyse du contexte linguistique apparaît ainsi indispensable pour nombre d'adjectifs moins marqués. Elle permet de repérer des indices qui expliquent la polarisation des adjectifs. Plusieurs mécanismes sur l'axiologisation et sur le changement de polarité des adjectifs ont été mis en évidence par Polanyi et Zaenen (2006). Par exemple, les connecteurs oppositifs (Ex. *mais, bien que*, etc.) relient généralement des adjectifs de polarité opposée comme dans *carte classique mais bonne* alors que *et* associera des éléments ayant une convergence de polarité (*hôtel classique et agréable*), etc.

Pour effectuer une étude de la langue axiologique dans les guides touristiques, dont l'objectif à plus long terme est l'étude des stéréotypes du discours touristique, nous avons réalisé une annotation des corpus. Cependant, comme nous l'avons vu, dans de nombreux cas, l'annotation ne peut pas

1. « 1 Qui retient l'attention, captive l'esprit. » (*Le Petit Robert* [en ligne], 2016)
2. « 1. Se dit d'un style et d'une attitude esthétique caractérisés par l'usage hétéroclite d'éléments démodés (→ **2. rétro**) ou populaires, considérés comme de mauvais goût par la culture établie et produits par l'économie industrielle. » (*Le Petit Robert* [en ligne], 2016)

être effectuée automatiquement, mais en prenant en compte le contexte, en suivant un ensemble de mécanismes comme ceux qui sont formulés par Polanyi et Zaenen (2006). Au-delà de ces mécanismes linguistiques, d'autres paramètres comme le facteur culturel, le genre ou la structure du texte doivent être pris en compte.

3. Méthodologie

Notre étude repose sur un corpus de 412 990 mots extraits de deux guides touristiques sur la Thaïlande : le Guide du Routard *Thaïlande* (2014) et le guide Gallimard *Thaïlande* (Chantraine et al., 2011). Lors d'un premier repérage, les deux guides nous sont apparus assez différents sur plusieurs plans : le public visé, le registre de langue, les thèmes abordés, etc. L'équipe de rédaction du Guide du Routard utilise en effet le style informel avec des expressions familières tandis que le Guide Gallimard préfère un style formel avec beaucoup d'informations culturelles.

Pour l'étude du lexique axiologique, nous nous appuyons principalement sur deux modèles théoriques : la notion d'axiologie selon la théorie des *subjectivèmes* élaborée depuis 1980 par Kerbrat - Orecchioni (2009) et le modèle de l'Attitude de l'*Appraisal* répandu dans le monde anglophone (Martin & White, 2005). Le modèle de Martin et White (2005) vise à catégoriser les unités lexicales qui partagent les mêmes caractéristiques. Parmi les sous-catégories proposées, c'est l'Attitude qui correspond le plus à notre travail concernant les adjectifs positifs/négatifs. Parmi les catégories proposées, le lexique de l'**Appréciation** sert à énoncer un jugement de valeur sur des cibles non conscientes telles que des lieux, des choses, etc. et se subdivise en plusieurs sous-types : **Réaction/Impact** qui regroupe les adjectifs attribuant aux cibles non conscientes l'affection du locuteur (Ex. *les rencontres indésirables*) ; **Réaction/Qualité** qui concerne les adjectifs décrivant la qualité des cibles (Ex. *bungalows très rudimentaires*) ; **Composition/Équilibre** qui exprime l'équilibre des caractéristiques des cibles, si elles sont bien assorties/discordantes (Ex. un *mélange harmonieux*) ; **Composition/Complexité** qui s'intéresse à la clarté, la simplicité, l'accessibilité, etc. (Ex. *la règle est simple*) ; et **Valorisation** qui concerne le jugement au niveau de la cognition (ex. *cadeau original*). Ces différentes catégories sont utilisées dans notre annotation et permettront de mettre en évidence les stéréotypes les plus prégnants.

Annotation

Le processus d'annotation est effectué semi-automatiquement, à l'aide du logiciel NooJ, développé par Max Silberztein (2014), à partir d'un lexique d'adjectifs axiologiques, dont les valeurs axiologiques se basent sur une référence dictionnairique, *le Petit Robert* [en ligne], à partir de leur définition [3]. À ce jour, notre dictionnaire d'adjectifs contient 1 220 adjectifs.

Dans notre lexique, nous intégrons trois types d'adjectifs axiologiques : (i) **les adjectifs intrinsèquement axiologiques** dont le sens central est polarisé d'après le Petit Robert (Ex. *mauvais, soigné*) (ii) **les adjectifs ayant une valeur axiologique dérivée** (Ex. *officiel*) lorsque le sens axiologique n'est pas central (iii) **les adjectifs axiologiques en contexte** qui prennent une valeur axiologique dans un contexte particulier du fait d'autres éléments linguistiques, du genre de discours touristique, ou d'autres facteurs. Par exemple, l'adjectif *ancien* devient positif en cooccurrence avec des cibles concernant le patrimoine culturel comme le temple, le site archéologique, etc.

Dans un deuxième temps, l'annotation semi-automatique est effectuée à l'aide de NooJ, à partir d'un graphe qui exploite le lexique d'adjectifs constitué à l'étape précédente. Plusieurs paramètres sont associés au traitement, comme nous le voyons dans l'exemple suivant, extrait du Guide du routard *Thaïlande* (2014).

Ex. 1 *Au fond, la base du temple originel et un très ancien et très vénéré bouddha de pierre.*

Sur le plan sémantique, l'exemple ci-dessus présente l'annotation de l'adjectif *ancien* qui est, par nature, un adjectif neutre. Nous le considérons ici toutefois comme un adjectif axiologique de polarité positive, car dans le contexte d'énonciation (*très ancien et très vénérable*, il énonce plutôt une attitude positive envers la cible du jugement (*bouddha*). Nous nous appuyons pour cela sur la cooccurrence avec *vénéré* qui induit cette valeur sémantique (la coordination avec et fonctionne principalement pour des convergences de polarité, cf. Polanyi et Zaenen (2006)). Au niveau de la catégorie sémantique, cet adjectif s'inscrit dans la catégorie Appréciation/Valorisation selon le modèle d'Attitude (Martin &

3. Du fait de la polysémie des adjectifs, cette étape est indispensable surtout pour les francophones non natives comme l'auteur (Jarukan Jitwonan).

Catégorie obligatoire Adjectif	*classique*
Type axiologique	Adjectif axiologique en contexte
Polarité	positive
Catégorie selon l'Appréciation	Appréciation/Valorisation
Cible évaluée	*Bouddha de pierre*
Type de cible	Site-temple-chose
Contexte grammatical	ADV A CONJC ADV A N
Fonction grammaticale	Épithète gauche
Catégorie facultative Adverbe d'intensité en cooccurrence	*très*
Conjonction en cooccurrence	*et*
Remarque	Contexte : *un très ancien et très vénéré*

Table 1 – Exemple de l'annotation d'adjectif axiologique

White, 2005). La fonction grammaticale de l'adjectif *ancien* est épithète. Quant à la catégorisation des cibles évaluées ayant pour but de mener une étude des stéréotypes, nous inscrivons la cible (*bouddha en pierre*) dans la catégorie « Site-temple-chose ».

Notre communication présentera les différents les différents cas de figures rencontrés pour les adjectifs non intrinsèquement axiologiques et les mécanismes linguistiques qui permettent d'identifier leur axiologisation.

Références bibliographiques

Chantraine, M., Demangeon, X., & Nee, K. (2011). *Thaïlande* (GALLIMARD-LOISIRS edition). Paris : Gallimard Loisirs.

Kerbrat-Orecchioni, C. (2009). *L'énonciation. De la subjectivité dans le langage* (4e éd.). Paris : Arnauld Collin.

Le Petit Robert [en ligne] (Version 5). (2016). (S.l.) : (s.n.). Repéré à http://pr.bvdep.com/login_.asp

Maingueneau, D. (2016). *Analyser les textes de communication* (Nouvelle éd. revue et mise à jour). Paris : Armand Colin. (22 cm. Bibliogr. p. 269-272. Index.).
Repéré à http://catalogue.bnf.fr/ark:/12148/cb45003546k

Martin, J., & White, P. (2005). *The Language of Evaluation : Appraisal in English.* New York : Palgrave.

Polanyi, L., & Zaenen, A. (2006). Contextual Valence Shifters. Dans J. G. Shanahan, Y. Qu, & J. Wiebe (Éd.), *Computing Attitude and Affect in Text : Theory and Applications* (pp. 1-10). (S.l.) : Springer Netherlands.
Repéré à http://link.springer.com/chapter/10.1007/1-4020-4102-0_1

Session 7.B.
Discours académique

A Comparative Study of Spatial Metaphors between Chinese and Western Academic Writing — take *"in"* and *"out"* as examples

Xinei Zhang
Germany Chemnitz Technology University

Abstract

This contribution uses a corpus-based approach to compare the usage of spatial metaphors *"in"* and *"out"* in non-native English academic writing by Chinese and western M.A. students quantitatively and qualitatively. The Chinese academic writing corpus contains more than 1.6 million words of Master' degree theses written by Chinese strudents in different branches of English language and literature between 2005 and 2014. The theoretical frames are the concept of embodiement philosophy and Conceptual Metaphr Theory.

In our quantitative study, the author makes an analysis of metaphorical expressions of *"in"* and *"out"* in one or more slots of constructions, thus yielding the contrast between Chinese and western learners in varous constructional senses.

In the qualitative study, many comparative exemples are demonstrated, which adduces the empirical evidence in support of different cognition s spatial metaphors between Chinese and western learners: despite the same schematicity on the basic level (literal meaning), learners have environment-specific "idiosyncrasies" in higher target domains. More preciously, western learners tend towards more abstract metaphorical usages than Chinese learners, while Chinese learners' texts show under-specification in the process of schematisation.

This contribution demonstrates that social-cultural factors (Chinese and western) greatly influence metaphorical usage, such as preposition choice. This confirms the philosophy of experientialism, according to which human basic perception comes from physical body, spatial relationship and movement action.

Key words: spatial Metaphor; collostructional analysis; *"in"* and *"out"*; image schema

References

Deignan, A. 2005b. *Metaphor and Corpus Linguistics.* Amsterdan & Philadelphia: John Benjamins.

Gibbs, R. W. Jr. 2006. *Embodiement and Cognitive Science.* Cambridge: Cambridge University Press.

Gries, S.T. 2001. A corpus-linguistic analysis of -ie and -ical adjectives. *ICAME Journal* 25:65-108.

Gries, S. T. 2003a. Testing sub-test: A collocational-overlap analysis of English -ie and -ical adjectives. *International Journal of Corpus Linguistics*8(1):31-61.

Gries, S. T. & Stephanowitsch, A. 2004. Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9(1):97-129.

Hanks, P. 2006. *Metaphoricity is gradable. Corpus-Based Approaches to Metaphor and Metonymy.* Berlin & New York: Mouton de Gruyter.

Johnson, M. 1987. *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason.* Chicago: The University Chicago Press.

Kövecses, Z. 2005. *Metaphor in Culture.* Cambridge: Cambridge University Press.

Lakoff, G. & Johnson, M. 1980. *Metaphors we live by.* Chicago & London: The University of Chicago Press.

Lakoff, G. 1987. *Women, Fire and Dangerous Things.* Chicago: The University of Chicago Press.

Lindstromberg, S. 2010. *English Prepositions Explained.* Amsterdam/Philadelphia: John Benjamins Publishing Press.

Schmied, J., Dheskali, J., Guo, Y., Zhang, X. (fc.). The Corpus of Chinese Academic Writting.

Tyler, A. & Evans, V. 2003. *The semantics of English Prepositions: Spatial Scenes Embodied Meanings and Cognition.* Cambridge: Cambridge University Press.

Lexical richness : Comparison of ELF, ESL, and L1 English oral academic presentations

Alla Zareva
Old Dominion University, USA
azarev@yahoo.fr

Research generally agrees that students' written production has received much more attention than their oral production (Read, 2000). As a result, we have a good number of studies that have researched extensively a variety of lexical aspects of English native speaking (L1) and English as a second language (ESL) students' writing but little research examining the lexical composition of their oral academic prose (Dang & Webb, 2014).

Student oral presentations are frequently used as an assessment assignment in higher education in North America as they reveal not only students' content knowledge about a certain topic but also how well students are able to communicate this knowledge to an audience of peers. One aspect of oral presentations which has been largely under-researched is their lexical richness, measured by the use of 1) vocabulary beyond the first 2K most frequent words, 2) academic vocabulary, 3) technical or discipline-specific vocabulary, 4) lexical density, and 5) lexical diversity (measured by MTLD [Measure of Textual Lexical Diversity (for a detailed review, see McCarthy & Jarvis, 2010]). Much less do we know about how English as a lingua franca (ELF), ESL and L1 speakers' oral academic performances compare on the lexical richness of these language users' prepared oral discourse. In light of some recent research on the use of English as a lingua franca for academic purposes, this line of investigation can potentially reveal the similarities (along with any differences) these three groups of proficient English language users share in their oral academic use of English.

The present study set out to examine the lexical richness of proficient ELF, ESL, and L1 graduate students' presentations, based on three small specialized corpora, each of which consists of over 60,000 words. All presentations were graded as high quality presentations by the respective instructors. Thus, the study aimed at finding out the lexical richness of good academic presentations, how it compared to findings about similar features of students' productively used vocabulary in academic writing, and what aspects of students' lexical uses may need to be addressed in oral communication courses, targeting the development of students' presentation skills at a graduate level of education.

The data were collected from graduate students' final project presentations (N = 90). The students were enrolled in various programs in Education, Applied Linguistics, International Studies, and the Humanities and were taking introductory level graduate courses. The presentations were compared on five features of lexical richness—i.e. 1) students' use of lower frequency vocabulary, 2) academic vocabulary, 3) specialized vocabulary, 4) lexical density, and 5) lexical diversity (measured by MTLD).

The analysis revealed more similarities than differences among the three groups of presenters in the lexical richness their academic presentations. It also revealed several main lexical differences between oral academic discourse and written academic prose and identified areas of similarities between oral academic and conversational prose in students' lexical choices. Finally, it uncovered few specific lexical areas that need to be addressed in English-based oral communication instruction for academic purposes to improve the smoothness of the presentations and increase students' awareness of the lexical acceptability in oral academic discourse.

A Case Study of Adjective-Noun Combination Used in Spoken Academic English

Fu-Ying Lin
Free University of Berlin
fuying88@gmail.com

1. Introduction

"Academic English" has long played a vital role in higher education, often considered a type of 'register' and worth teaching with a specific case of English for Special Purposes (cf. Brisk and Jeffries 2008 and Bailey 2012 for overviews and Benesch 2001 for a critical appraisal). Academic English is largely treated as a written register with occasional exceptions, which often tend to be tiny portion of language materials ; for example, spoken materials used in the New Academic Word list (NAWL 1.0, cf. Browne, C. et.al 2013) is merely 1%, which is unlikely to have a noticeable effect on the word selection process. Although theoretical researchers are more inclusive of spoken language on average, researchers like Biber and Grey also equate Academic English with written academic prose in their investigation (Biber and Grey 2016). Where spoken language is included, the definition of "Academic English" is often too broad, covering any type of classroom discourse, for instance, in the Corpus of English as Lingua Franca in Academic Settings (ELFA 2008) or in the Michigan Corpus of Spoken Academic English (MICASE, Simpson et al 2002).

With the increasing popularity of Massive Open Online Courses (MOOCs) involved in higher education, the visibility of the spoken contents in Academic English is worth of more severe manifestation (see Bailey 2012 :7 ; Anthony 2015). This paper will investigate the frequency of the Adjective-Noun combination manipulated on Spoken Academic English by pointing out its quantitative frequency in the specifically-conducted MOOC corpus (compiled by me) with the colloquial corpus, Corpus of Contemporary American (COCA), as its reference.

2. Background

MOOC represents the course that are open to large amounts of learners (massive), freely accessible to students world-wide(open), accessible via the Internet (online), emerging in 2008, now a new extension of higher education in recent years (Fini 2009 ; Kennedy 2014). The enormous number of platforms, organizations and institutes joining the MOOC movement with offerings of their own has made the year 2012 the year of MOOC (Pappano 2012). At least 26 providers are covered in MOOCs, three of which can be considered major providers—Coursera, Edx and Udacity (Ha 2014). As a novel learning opportunity/tool, MOOCs bring about more availability to potential students all over the world. However, the good command of academic English, especially the Spoken Academic English, is demanded.

Known for being different from general English (Coxhead and Nation 2001 ; Schmitt 2000 ; Xue and Nation 1984), Academic English features its own professional vocabulary (not limited to technical terminology), grammatically distinct with its complex, informationally-dense structures (Biber et.al 1999 ; Biber 2006). Biber and Gray's recent research (2016) even challenges some of the stereotypical views of this register, but still falls into an implicit equation of academic English with written English. Spoken corpora like the 1.8-million-word Michigan Corpus of Spoken Academic English (MICASE) corpus (Simpson et al 2002), the British Academic Spoken Corpus (BASE) (Nesi and Thompson 2001), the Corpus of English as Lingua Franca in Academic Setting (ELFA 2008), and the 2-million-word Hong Kong Corpus of Spoken English (HKCSE)(Warren, Chen 2004), have collected language data from various academic settings like classrooms, meetings, seminars, conferences and job interviews, whose content ranges have made the definition of Spoken Academic English too expensive to be used.

Therefore, I will define Spoken Academic English more narrowly as the spoken language used by academics talking about their subject in academic settings, i.e. including talks by and discussion amongst researchers at academic conferences as well as lectures by instructors in physical or virtual university classrooms, excluding academics using language in non-academic university settings such

as faculty meetings or classroom language by students. The language of MOOCs clearly falls under this definition. Since the research to date tends to focus on (written) academic English, and limited research only characterize Spoken Academic English as an (emerging) register, the Spoken Academic English found in the context of MOOCs is worthwhile, plausibly argued as an identifiable subgenre "EMP" (English for MOOC Purposes) (see Anthony 2015).

3. Methodology

I have specifically constructed a corpus, the MOOC corpus, consisting of 93 lecture transcripts from MOOCs, amounting to 8,716,104 tokens, in five broad subjects represented offered by the major MOOC providers (Coursera, EdX, Udacity and Futurelearn).The subjects are those widely agreed upon in studies of Academic English, represented in such large corpus as the British Academic Written English Corpus(BAWE)(Garnder & Nesi 2012, 2013). In order to identify the unique properties of Spoken Academic English, a corpus of general spoken English (the spoken part of COCA, the Corpus of Contemporary American English)(Davies 2010) will be treated as the reference corpora.

4. Results & conclusion

Academic English is assumed to be characterized both by domain-specific vocabulary and by vocabulary also found in colloquial language but used with specific technical meanings in academic settings. Terminological vocabulary may consist of single words, but more often it consists of multi-word expressions of sequences like Adjective-Noun or Noun-Preposition-Noun (Justeson and Katz 1995). Comparing the frequency of, in this paper, Adjective-Noun combinations in the MOOC corpus against their frequency in the spoken part of the COCA and extracting combinations that occur significantly more frequently in the former than in the latter should thus allow us to identify academic terminology.

Table 1 shows the top 15 adjective-noun combinations that are significantly more frequent in the MOOC data than in the spoken part of the COCA with their observed and expected frequencies, the log-likelihood value and the information whether they occur in both corpora or only in the MOOC data

		MOOC data		COCA spoken data			
	WORD	Observed	Expected	Observed	Expected	LL	Shared
1	*random variable*	1419	320.8	0	1098.2	4222.74	No
2	*random variables*	834	188.6	0	645.4	2481.05	No
3	*special theory*	722	163.9	3	561.1	2110.35	Yes
4	*fair use*	465	106.7	7	365.3	1313.77	Yes
5	*next video*	467	109.2	16	373.8	1256.68	Yes
6	*next lecture*	348	78.9	1	270.1	1021.78	Yes
7	*differential equation*	343	77.8	1	266.2	1006.94	Yes
8	*gradient descent*	338	76.4	0	261.6	1005.23	No
9	*absolute value*	315	72.4	5	247.6	887.87	Yes
10	*initial state*	298	67.8	2	232.2	863.24	Yes
11	*solar system*	529	175.9	249	602.1	725.50	Yes
12	*white dwarf*	240	54.5	1	186.5	701.28	Yes
13	*kinetic energy*	255	60.1	11	205.9	672.36	Yes
14	*straight line*	371	197.6	105	368.4	654.86	Yes
15	*conditional expectation*	219	49.0	0	169.5	651.27	No

TABLE 1 – Ad-N combinations significantly more frequent in the MOOC corpus compared to COCA

The approach clearly yields mostly scientific terminology, either in a narrow (domain-specific) sense (like random variable(s), differential equation, gradient descent, kinetic energy) or in a general sense (fair use, absolute value, initial state, solar system). The former tends to be restricted to the MOOC data, while the latter tends to occur in the general language too, but have a special technical meaning in academic English.

References

Ackermann, K., De Jong, J.H.A.L., Kilgarriff, A. & Tugwell, D. (2010). The Pearson International Corpus of Academic English (PICAE)

Anthony, L. (2015). The Changing Role and Importance of ESP in Asia. English as a Global Language Education (EaGLE) Journal, 1(1) : 01-21. DOI :10.6294EaGLE.2015.0101.01

Coxhead, A., & Nation, I.S.P. (2001). *The specialized vocabulary of English for academic purposes.* In. J. Flowerdew & M. Peacock (Eds.), Research perspectives on English for academic purpose (pp.252-267). Cambridge : Cambridge University Press.

Bailey, L. (2012). Academic English. In James Banks (ed.), *Encyclopedia of Diversity in Education*, 4–9. Thousand Oaks, CA : SAGE.

Benesch, S. (2001). *Critical English for academic purposes : theory, politics, and practice.* Mahwah, N.J : L. Erlbaum Associates.

Biber, D., Stig J., Geoffrey N. L., Susan C. & Edward F. (1999). *Longman grammar of spoken and written English.* Harlow : Longman

Biber, D. & Gray, B. (2016). *Grammatical Complexity in Academic English : Linguistic Challenge in Writing.* Cambridge : CUP

Brisk, M. E. & Julian J. (2008). *Academic English. In Josué González (ed.), Encyclopedia of Bilingual Education*, 1–4. Thousand Oaks, CA : SAGE

Browne.C., Culligan, B. & Phillips, J. (2013). The New Academic Word List. Retried from http://www.newgeneralservicelist.org

Cheng, W., Greaves, C. and Warren M. (2005). The creation of a prosodically transcribed intercultural corpus : The Hong Kong Corpus of Spoken English (prosodic). International Computer Archive of Modern English (ICAME) 29. 5-26.

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* 25(4). 447–464. doi :10.1093/llc/fqq018

ELFA. (2008). The Corpus of English as a Lingua Franca in Academic Settings. Director : Anna Mauranen. http://www.helsinki.fi/elfa/elfacorpus

Fini, A. (2009). The technological dimension of a massive open online course : The Case of the CCK08 course tools. The International Review of Research in Open and Distance Learning, 10(5). Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/643/1410

Gardner, S. & Nesi, H. (2013). A classification of genre families in university student writing. *Applied Linguistics*, 34 (1) 1-29.

Ha, T.H. 2014. *MOOCs By the Numbers : Where Are We Now ?* in the blog of ideas.ted.com, retrieved from http://ideas.ted.com/moocs-by-the-numbers-where-are-we-now/

Justeson, J. S. & S. M. Katz (1995) Technical terminology : some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* 1 : 9–27.

Kennedy, J. (2014). Characteristics of Massive Open Online Courses (MOOCs) : A Research Review. Journal of Interactive Online Learning, 13.

Nation, P. & Xue, G. (1984). A university word list. Language Learning and Communication, 3(2), 215-229.

Pappano, L. (2012). The year of the MOOC. New York Times. Retrieved from http://www.nytimes.com/2012/11/04/education/edlife/massive-open-online-courses-are-multiplying-at-a-rapid-pace.html?pagewanted=all&_r=1&

Schmitt, N. (2000). *Vocabulary in Language Teaching.* Cambridge : Cambridge University Press.

Simpson R.C., S. L. Brigss, J. Ovens and J. M. Swales. (2002). T*he Michigan Corpus of Academic Spoken English*, Ann Arbor, MI : The Regents of the University of Michigan.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing.* Manchester.

Schmid, H. (1995). *Improvements in part-of-speech tagging with an application to German.* Proceedings of the ACL SIGDAT-Workshop. Dublin.

Thompson, P. and Nesi, H. (2001). The British Academic Spoken English (BASE) Corpus Project. Language Teaching Research (3), 263-26

Posters

Étude terminologique des verbes d'un corpus spécialisé : le cas de la chimie en arabe

Baïan Albeiriss
Laboratoire ICAR (Interactions, Corpus, Apprentissage, Représentations),
UMR 5191-CNRS Université Lyon 2 - ENS de Lyon
15 parvis René Descartes BP-7000, 69342 Lyon cedex 07, France
baian.albeiriss@univ-lyon2.fr

1. Introduction

Avec le développement de l'informatique, les corpus électroniques apparaissent, explosent et deviennent le support d'information des chercheurs en favorisant et facilitant la constitution de corpus de grande taille, condition nécessaire (Sinclair, 1991), mais non suffisante. À cela, l'apparition de bases de données, de thésaurus, d'ontologies... et face aux besoins des entreprises qui doivent gérer un volume de données augmentant sans cesse, avec une durée de vie de l'information de plus en plus courte, pour les rendre accessibles le plus rapidement possible, les linguistes en général, les terminologues en particulier (Condamines, 2005) sont au cœur des problématiques du TAL, de la linguistique de corpus et de l'ingénierie linguistique.

Notre travail de recherche a pour objectif la conception d'un outil d'extraction automatique des termes de la chimie en arabe ; mais réaliser un tel outil de traitement ne peut se faire sans analyse linguistique approfondie basée sur des corpus. Dans cet objectif, nous avons constitué un corpus de textes à partir de ressources spécialisées et nous l'avons traité à l'aide d'outils informatiques afin de permettre la description sémantique des termes ainsi que la description des mécanismes de leur formation. Cette étude terminologique nous permettra de construire une grammaire d'identification des unités terminologiques simples et complexes ; les règles de cette grammaire d'identification pourront être implémentées par la suite dans un système d'analyse morpho-syntaxique de l'arabe afin d'extraire les termes du domaine.

Dans cette présentation, nous décrivons les verbes de la chimie en arabe, en étudiant leur structure argumentale (L'Homme, 2012), notamment en définissant les distinctions sémantiques et en annotant les structures syntaxiques.

2. Identification des verbes

Les termes sont majoritairement des noms et des adjectifs (L'Homme, 1998) ; mais, les verbes de la langue générale peuvent être aussi considérés comme des termes dans un domaine de spécialité en général, en chimie en particulier. En effet, le verbe a longtemps été mis de côté au profit du nom en terminologie, cela étant justifié notamment par la place accordée aux objets et à leurs dénominations (Rey, 1979). Ce n'est que récemment que le verbe s'est vu accorder une place dans les travaux terminologiques (Pimentel, 2012), le verbe pouvant notamment soulever un problème de compréhension.

2.1. Constitution du corpus

À partir des ressources spécialisées, nous constituons un corpus de textes, auquel nous ne fixons pas, à terme, de limite de taille et nous choisissons de couvrir tout le domaine de la chimie quels que soient les thèmes et les sujets, afin de prétendre à l'exhaustivité du domaine et non à sa représentativité, en limitant notre recherche au monde universitaire, englobant les travaux des enseignants et des étudiants, afin de garantir un niveau de spécialisation et de vulgarisation des textes mais aussi leur fiabilité ainsi que leur authenticité. Le corpus recueilli est modeste et n'est constitué que d'une petite centaine de milliers de mots ; il est issu du département de chimie de l'université de Constantine 1 en Algérie.

2.2. Dépouillement du corpus

La sollicitation des corpus, notamment des méga-corpus, a suscité le développement des outils informatiques destinés à leur analyse, tels que les concordanciers, les analyseurs ou les extracteurs, atteignant un nombre conséquent d'outils mis à disposition pour les langues comme l'anglais, le français,

l'espagnol... mais l'arabe voit ce foisonnement d'outils s'effondrer tant par sa graphie agglutinante que par la complexité de la reconnaissance optique de ces caractères (Zghibi, 2002). Bien qu'elle soit parlée par presque 300 millions de personnes, l'arabe ne dispose pas d'assez de ressources, notamment des ressources gratuites (Meftouh and al., 2007) et le manque de coordination de méthodologies et de standards entre équipes ne favorisent pas la recherche qui n'a commencé que récemment pour cette langue. De ce fait, si certains s'affranchissent de ce problème en privilégiant le traitement manuel, nous pensons qu'une automatisation des traitements peut être possible en arabe et nous explorons les différents outils informatiques proposés dans la littérature en les combinant et/ou en les juxtaposant.

La sélection des verbes, est soumise aux critères « ...selon lesquels une unité lexicale prédicative peut être un terme si : ses actants et ses dérivés morphologiques sont des termes et les dérivés morphologiques s'apparentent sémantiquement au terme, l'unité lexicale entretient des relations paradigmatiques avec d'autres termes ». (Pimentel, 2012). Ces critères reposent sur le lien qui peut être établi avec des termes de nature nominale et la classe sémantique des verbes, impliquant que la participation des verbes à la transmission de connaissances spécialisées passe par leur lien plus ou moins étroit avec des noms définis comme termes (L'Homme, 2012).

Chaque verbe, par exemple « جفّف = *jaffafa* = sécher », est indexé en indiquant sa forme, « forme II (فَعَّلَ = *faala*) », son nombre d'occurrence, « 10 », et son contexte, « يجفف الإيثر فوق $CaCl_2$ مدة 24 ساعة = *yujaffafu al'îṯir fawqa CaCal2 muda 24 sâ'a.* = L'éther est séché sous $CaCl_2$ pendant 24 heures », à l'aide des logiciels AntConc, Kawâkib et Xerox.

Nous identifions une trentaine de verbes comme des termes du domaine de la chimie.

3. Description des verbes

Dans la description des verbes de la chimie en arabe, nous analysons chaque terme et nous précisons les concordances, les relations sémantiques et les éléments de la structure argumentale ; ces informations permettent de représenter les arguments du verbe dans la situation évoquée. Chaque verbe se voit attribué une fiche terminologique, contenant ces informations ainsi que son lemme, sa forme et ses contextes.

3.1. Classification des verbes

La classification permet d'organiser les verbes selon leur niveau de spécialisation (Gross, 2008), afin d'organiser les concepts du domaine, pour d'en faciliter l'accès et l'étude (Pavel et Nolet, 2001). Des verbes de sens généraux, à savoir les verbes énonciateurs permettant d'articuler le discours (comme شكّل = *šakkala* = constituer), aux verbes de sens spécialisés, à savoir les verbes dont l'usage est spécifique au domaine de la chimie (comme أكسد = *'aksada* = oxyder), en passant par les verbes polysémiques dont au moins un sens est spécialisé (comme قطّر = *qaṭṭara* = filtrer), cette classification a pour but d'exclure les sens généraux et d'isoler les sens spécialisés (Lerat, 2002).

3.2. Structure argumentale

La structure argumentale se présente comme une sorte de paraphrase qui précise le nombre d'arguments et décrit leur position par rapport au terme (Tellier, 2008). Pour chaque argument identifié, une étiquette est attribuée ; par exemple, « يتشكل ناتج التفاعل الألدولي = *yatašakalu nâtij altafâ'ul al'uldûlî* = il se forme le produit de la réaction aldolique » possède deux arguments. La création d'étiquettes, pour notre exemple « Constituant Chimique », permet de décrire les groupes d'arguments qui partagent des caractéristiques sémantiques communes et qui servent de génériques pour les formes spécifiques que nous avons relevées dans le corpus. Les étiquettes permettent donc de relier la structure du verbe, dégagée à partir des textes analysés, à l'information sémantique contenue dans le corpus. Elles ont été établies par l'examen et la comparaison des arguments repérés dans l'environnement des verbes spécialisés.

3.3. Structure des syntaxes terminologiques

La structure des syntaxes terminologiques indique les éléments syntaxiquement nécessaires pour décrire les verbes. En effet, « elle indique plus précisément dans quel ordre les actants s'articulent dans la phrase et, s'il y a lieu, la ou les prépositions privilégiées » (L'Homme, 1998). La complexité de la description des constructions syntaxiques varie selon les possibilités d'inverser l'ordre des arguments sans affecter le sens du verbe et de sélectionner diverses prépositions. Nous avons également remarqué que la sélection de la préposition est en lien avec les étiquettes des arguments d'un verbe.

4. Discussion

A partir de l'identification et de la description des verbes, nous modélisons les verbes de la chimie en arabe, en prenant en compte leur structure argumentale ; par exemple, le verbe « قطّر = *qaṭṭara* = distiller » présente trois modélisations : « Verbe + prép + Constituant Chimique », « Verbe + Constituant Chimique + prép + Constituant Chimique » et « Verbe + Constituant Chimique + prép + Constituant Chimique prép + Grandeur ».

D'autre part, nous observons que certaines nominalisations sont beaucoup plus fréquentes que le verbe correspondant. En effet, l'emploi des noms et des adjectifs est prédominant dans les domaines de spécialité. Nous avons ainsi étudié plusieurs dizaines de couples nom/verbe comme le couple « تحضير / حضّر = *ḥaḍḍara* / *taḥḍîr* = préparer / préparation ». Pour repérer les noms morphologiquement reliés à des verbes, nous avons élargi notre zone de recherche en ciblant des couples comme sécher / séchage. Par abus de langage, nous désignerons par « nom d'action » les noms appartenant à ces couples (Bourigault & Condamines 1999).

De plus, nous remarquons que les formes actives et passives des verbes arabes de notre corpus ne se distinguent pas en l'absence de voyellation par un outil de traitement automatique de la langue. En effet, le verbe « قدّر = *kadarra* = estimer », en l'absence de voyelles peut se lire à la voix active « يقدر = *yuqadirru* = il estime » ou à la voix passive « يقدر = *yuqadarru* = il est estimé ». Cela a pour conséquence la modification de la structure argumentale du verbe notamment l'ordre de ses arguments.

À cela, nous repérons l'importance des prépositions dans la structure argumentale des verbes. En effet, le verbe « جفّف = *jaffafa* = sécher » se décrit par un premier argument indiquant le composé chimique à sécher suivi d'une préposition introduisant un second argument expliquant avec quel composé chimique l'opération se réalise ; il peut être aussi accompagné d'une seconde préposition informant de la durée du séchage. Ces informations sont prises en compte dans la modélisation de la structure argumentale des verbes.

5. Conclusion

A partir d'un corpus relativement modeste, nous avons recueilli les différentes structures argumentales des verbes de la chimie ainsi qu'une vaste variété de patrons de termes et d'indications relatives aux termes eux-mêmes, en partant de l'identification des termes jusqu'à leur classification, en passant par leur formation. Les caractéristiques terminologiques de ces verbes ont permis de les modéliser ce qui rend possible à présent la construction d'une grammaire d'identification des unités terminologiques simples et complexes. Par la suite, les règles de cette grammaire d'identification pourront être implémentées dans un système d'analyse morpho-syntaxique de l'arabe afin d'extraire les termes du domaine, notre travail étant la conception d'un extracteur morpho-syntaxique robuste pour la fouille de textes dont l'objectif final étant la recherche d'information.

Références bibliographiques

Arbach, A. & Ali, S. (2013). Aspects théoriques et méthodologiques de la représentativité des corpus, *Corela*, HS-13.

Cabré, M. T. (2008). Constituer un corpus de textes de spécialité : bilan et perspectives. *Les Cahiers du Cel.* Paris, UFR d'Études Interculturelles de Langues Appliquées, 37-56.

Condamines, A. (1999). Alternance nom/verbe : explorations en corpus spécialisé. B. Victorri et J. François (eds) : *Sémantique du lexique verbal, Actes de l'atelier de Caen,* Cahiers de l'Elsap, 41-48.

Condamines, A. (2005). Linguistique de corpus et terminologie, *Langages,* 157, 36-47.

Gross, G. (1994). Classes d'objets et description des verbes. *Langages,* 115. 15-30

Gross, G. (2008). Les classes d'objets. *Lalies,* 28, 111-165.

L'Homme, M.-C. (1998). Le statut du verbe en langue de spécialité et sa description lexicographique. *Cahiers de lexicologie* 73(2), 61-84.

L'Homme, M.-C. (2012). Le verbe terminologique un portrait de travaux récent. *Congrès Mondial de Linguistique Française-CMLF*, 93-107.

Lerat, P. (2002). Qu'est-ce que le verbe spécialisé ? Le cas du droit. *Cahiers de Lexicologie,* 80, 201-211.

Meftouh K., Smaïli K. & Laskri M. T. (2007). Constitution d'un corpus de la langue Arabe à partir du Web. *Colloque International sur le Traitement Automatique de la Langue Arabe* - CITALA'07, Rabat, Maroc.

Pavel S. et Nolet D. (2001). *Précis de terminologie,* Canada, ISBN 0-660-61616-5, No de cat. S53-28.

Pimentel, J. (2012). Description de verbes juridiques au moyen de la sémantique des cadres, *Terminologie & Ontologie : Théories et applications (Toth 2011)*, Annecy 2011.

Pimentel, J. et L'Homme, M.C. (2011). Annotation syntaxico-sémantique de contextes spécialisés : application à la terminographie bilingue. In van Campenhoudt, M., T. Lino et R. Costa (éd.). *Passeurs de mots, passeurs d'espoir : lexicologie, terminologie et traduction face au défi de la diversité,* Paris : Édition des archives contemporaines/Agence universitaire de la francophonie, pp. 651-670.

Rey, A. (1979). *La terminologie : noms et notions.* Coll. « Que sais-je ? », Paris : Presses universitaires de France.

Tellier, C. (2008). *Verbes spécialisés en corpus médical : une méthode de description pour la rédaction d'articles terminographiques,* Travail dirigé présenté au Département de linguistique et de traduction, Université de Montréal.

Sinclair, J. (1991). *Corpus, concordance, collocations,* Oxford, Oxford University Press.

Zghibi, R. (2002). Le codage informatique de l'écriture arabe : d'ASMO 449 à Unicode et ISO/CEI 10646, *Document numérique,* 6 (3), 155-182.

Utilisation des corpus pour la description de l'idiome *faire le malin*

Elena Berthemet
Chercheur associé au Centre de Linguistique en Sorbonne
elenaberthemet@gmail.com

1. Introduction

Il n'est plus à prouver que les corpus présentent de nombreux avantages et sont un outil indispensable au service de la linguistique théorique et appliquée. En lexicographie, grâce à une base empirique étendue, les corpus permettent aux spécialistes de peaufiner la description des mots dans les dictionnaires. L'objectif de la présente communication est de promouvoir l'utilisation des corpus multimodaux dans l'étude des expressions idiomatiques. Par conséquent, les corpus seront considérés du point de vue de leur exploitation et non de la constitution. Cependant, les résultats pourront être utilisés par des chercheurs constituants des corpus idiomatiques.

2. Problématique

L'étude des expressions polylexicales remonte au *Traité de stylistique française* de Charles Bally. Aujourd'hui, les travaux dédiés à la phraséologie sont variés et se développent dans plusieurs directions (A. Baranov/D. Dobrovol'skij 2008, 2013, H. Burger/D. Dobrovol'skij/P. Kühn/N. R. Norrick 2007, D. Dobrovol'skij/E. Piirainen 2005, A. Cowie 1998, I. Gonzalez-Rey 2002, S. Granger/F. Meunier 2008, F. Grossmann/A. Tutin 2003, R. Moon 1998, J. Sinclair 1991, A. Wray 2008). En France, l'accent est essentiellement mis sur les *collocations*, c'est-à-dire, les unités à sens compositionnel (P. Blumenthal/F. Hausmann, F. Grossman/A. Tutin, I. Mel'čuk/A. Polguère, G. Williams/S. Vessier). Nous tenterons d'étendre la recherche aux unités à sens non-compositionnel, appelées *expressions idiomatiques*. Notre travail part du constat suivant : la plupart des dictionnaires français, qu'ils soient généraux ou spécialisés, délaissent la description des unités en question.

Pour montrer en quoi les corpus peuvent aider lors de la description des expressions idiomatiques, nous allons prendre comme exemple l'expression *faire le malin*. L'idée de travailler autour de cette unité est venue d'une expérience menée auprès d'étudiants apprenant le français langue étrangère du niveau B1+. Le résultat de l'expérience était inattendu : cette unité a posé des problèmes non seulement d'utilisation mais aussi de compréhension. Une question s'était alors posée : Pourquoi ni les exemples scrupuleusement sélectionnés par l'enseignante à partir de grands corpus, ni les dictionnaires n'ont été utiles pour la compréhension du sens de cette unité ?

3. Hypothèse

Nous partons de l'hypothèse que les corpus permettent une description fine et exhaustive satisfaisant à la fois la compréhension et la production du lexique. Cependant, les expressions idiomatiques, plus encore que les lexies simples, sont difficiles à décrire. Tout d'abord, parce qu'elles sont dotées d'une *structure sémantique complexe* où des phénomènes lexicaux, syntaxiques et pragmatiques interagissent. Ensuite, parce qu'elles contiennent des *images* faisant partie de la *charge culturelle partagée* (Galisson 1991) et témoignant des valeurs d'une communauté langagière.

Enfin, il est important de souligner que les sujets parlants utilisent les idiomes dans le discours non seulement pour communiquer des informations descriptives, mais aussi pour exprimer leurs *sentiments*, leur attitude généralement négative ou parfois positive envers le sujet dont ils parlent. Suivant John Sinclair, nous retiendrons le terme *prosodie sémantique* pour parler de ce phénomène : « Souvent, l'utilisation d'un mot dans un cotexte particulier implique un sens supplémentaire de nature émotive ou attitudinale » (« Often the use of a word in a particular cotext carries extra meaning of an emotive or attitudinal nature » (Sinclair 2003 : 117).

Par conséquent, les corpus multimodaux, comme le *Corpus de référence du français* CRFC, permettraient de les étudier dans des discours réels et seraient un outil le mieux approprié pour la description de ces unités. En effet, les corpus multimodaux permettent d'entendre l'intonation et les pauses, ainsi que de visualiser les gestes et les mimiques qui participent à la construction du sens. Cependant, le

CRFC n'étant pas encore disponible pour le grand public, nous nous tournons vers les corpus présents sur le marché.

4. Cadre théorique et méthodologie

Nous nous appuierons sur les recherches de l'*Ecole sémantique de Moscou* et de Baranov et Dobrovol'skij (2008, 2013) dont nous retiendrons les idées suivantes : le sens lexical doit être étudié en contexte ; la compréhension n'est pas universelle ; les expressions idiomatiques ont une structure sémantique et conceptuelle complexe ; la description de ces unités doit être intégrale.

La méthode que nous avons employée consiste à prendre comme point de départ quelques définitions extraites des principaux dictionnaires disponibles à l'heure actuelle. Les dictionnaires unilingues français suivants seront considérés : tout d'abord, les généraux (dont le TLFi), ensuite, les spécialisés (dont Rey, Chantreau 1997).

Après avoir analysé les définitions de ces dictionnaires et avoir fait quelques conclusions temporaires, nous procéderons à la recherche d'un corpus proposant cette expression idiomatique. Notre objectif sera de trouver un corpus suffisamment riche et fournissant au moins une centaine d'extraits de textes. Les corpus disponibles à l'adresse suivante seront examinés : https://the.sketchengine.co.uk/. Cependant, les corpus, malgré toute leur richesse, sont parfois pauvres en expressions idiomatiques. C'est pour cela que nous nous contenterons du corpus frTenTen qui propose près de dix milliards de mots.

Lors de la phase principale, les contextes fournis par le corpus unilingue seront analysés. Tout d'abord, nous tenterons de distinguer les principaux types de contextes qui serviront à déterminer les caractéristiques sémantiques principales de l'unité en question. Ensuite, s'appuyant sur les types de contextes, nous écrirons une proposition de définition. Il est fort probable qu'un certain nombre de contextes soient écartés de la typologie proposée : soit parce qu'ils se trouvent à la limite des deux types, soit parce que leurs contextes présentent des bruits empêchant la compréhension et, par conséquent, ne permettent pas l'attribution d'un exemple à un type de contexte.

Nous tenterons ensuite de considérer des exemples provenant du *Nacional'nyj korpus russkogo âzyka* Corpus national de la langue russe (NKRÂ) avec, comme objectif, de voir si les corpus parallèles permettent de relever d'autres composantes de sens et, par conséquent, d'affiner l'analyse sémantique. La dernière étape sera de comparer les sens proposés dans les dictionnaires et ceux qui ont résulté des contextes proposés par le corpus.

Enfin, nous verrons dans quelle mesure les corpus dont il est question dans la présente étude permettent d'obtenir des informations à propos des caractéristiques prosodiques, de la gestuelle et des mimiques qui accompagnent souvent les idiomes.

5. Conclusion

Au terme de cette étude, il apparaît que la définition issue des dictionnaires français traditionnels ne permet pas de capturer avec précision l'ensemble des sens d'une expression. Peut-on dire que les corpus analysés aient résolu tous les problèmes de la description des idiomes ? Il nous semble que, pour ce qui est des corpus français disponibles à l'heure actuelle, la réponse soit négative.

En effet, la plupart expressions idiomatiques appartiennent à la communication orale non-officielle. Or, les corpus utilisés, écrits, ont fait surgir des problèmes sous-jacents concernant l'intonation, les mimiques et la gestuelle dont les expressions idiomatiques sont souvent dotées. Par conséquent, il serait intéressant de les étudier dans des corpus multimodaux, comme le *Corpus de référence du français* CRFC prochainement disponible.

De plus, les corpus explorés dans le cadre de la présente étude ne disent rien sur son ancrage culturel ni sur la situation extra-linguistique où l'expression idiomatique est employée. Certains soutiendrons qu'il s'agit d'informations secondaires, et qu'il est légitime de les négliger dans la description, mais, dans un souci de précision, nous nous penchons vers l'idée qu'aucune information distinctive ne doit être négligée dans la description.

De par leur étendue, les corpus présentent un gisement linguistique précieux pour tout spécialiste de langue. Cependant, les données ne parlent pas, et, comme l'affirme Henri Béjoint (2007), « il est

clair que le corpus ne peut pas tout faire seul ». Tout d'abord, parce que les données ne peuvent pas être prises sans avoir été triées et ensuite, parce que une introspection est nécessaire pour comprendre ce qu'elles signifient pour la théorie de la langue. Par conséquent, « la déduction et ... [le] pouvoir de synthèse » sont indispensables de la part du chercheur (Béjoint 2007).

Références bibliographiques

Baranov A. N., Dobrovol'skij D. O. (2008) *Aspekty teorii frazeologii.* Moskva : Znak.

Baranov A. N., Dobrovol'skij D. O. (2013) *Osnovy frazeologii.* Moskva : Flinta Nauka.

Béjoint H. (2007) Informatique et lexicographie de corpus : les nouveaux dictionnaires. *Revue française de linguistique appliquée,* Vol. XII, p. 7-23.

Burger H., Dobrovol'skij D., Kühn P., Norrick N. R. (2007) Phraseology. *An International Handbook of Contemporary Research.* Berlin, New York : Walter de Gruyter.

Blumenthal P., Hausmann F. J. (éd.) (2006) *Collocations, corpus, dictionnaires.* Paris : Larousse.

Cowie A. P. (ed.) (1998) *Phraseology. Theory, Analysis, and Applications.* Oxford : Clarendon Press.

Dobrovol'skij D., Piirainen E. (2005) *Figurative language. Cross-Cultural and Cross-Linguistic Perspectives.* Amsterdam : Elseveir.

Galisson R. (1991) *De la langue à la culture par les mots.* Paris : CLE International.

Gonzalez-Rey I. (2002) *La phraséologie du français.* Toulouse : Presses Universitaires du Mirail.

Granger S., Meunier F. (2008) *Phraseology. An interdisciplinary perspective.* Amsterdam, Philadelphia : John Benjamins Publishing Company.

Grossmann F., Tutin A. (2003) *Les collocations : analyse et traitement.* Amsterdam : De Werelt.

Mel'čuk I., Polguère A. (2007) *Lexique actif du français : l'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français.* Bruxelles : De Boeck.

Moon R. (1998) *Fixed expressions and idioms in English.* Oxford : Clarendon press.

Rey A., Chantreau S. (1997) *Dictionnaire des expressions et locutions.* Paris : Dictionnaires Le Robert.

Sinclair J. (1991) *Corpus, concordance, and collocation.* Oxford : Oxford University Press.

Sinclair J. (2003) *Reading concordances : An introduction.* London : Longman.

Trésor de la Langue Française informatisé, http://atilf.atilf.fr/tlf.htm

Williams G., Vessier S. (2004) *Proceedings of the eleventh EURALEX international congress, EURALEX 2004.* Lorient, France. Lorient : Université de Bretagne-Sud, Faculté des lettres et des sciences humaines.

Wray A. (2008) *Formulaic language : Pushing the boundaries.* Oxford : Oxford University Press.

Analyse sémantique du discours écologique relatifs au 雾霾 (wù maí), «brouillard de pollution» en Chine

Qinran Dang et Mathieu Valette
INALCO, ERTIM
qinran.dang@inalco.fr, mvalette@inalco.fr
mailto :mvalette@inalco.fr

1. Introduction

Depuis 2008, la pollution préoccupe sérieusement la population chinoise. En 2011, alors que la pollution de l'air prend une ampleur inédite, une dénomination dédiée - 雾霾 (*wù maí, i.e.* « brouillard de pollution » - fait son apparition dans les médias chinois et circule depuis sur les réseaux sociaux. En tant que problème environnemental, le *wù maí* revêt une dimension sociale et économique et devient omniprésent sur les sites, la presse et les différents réseaux sociaux. Dans le cadre de la présente recherche, nous procéderons à l'étude contrastive du champ thématique du *wù maí* dans trois corpus correspondant à trois types de discours (institutionnel, médiatique informel et médiatique institutionnel). Nous articulons les études lexicales et analyses textométriques (au moyen de Lexico 5). Notre hypothèse est que les trois corpus utilisent systématiquement des lexiques spécifiques [1] pour exprimer un thème similaire (en l'occurrence les causes de *wù maí*), et nous soupçonnons le choix des lexiques d'être lié au type du corpus auquel ils sont attachés. Finalement, notre objectif est de montrer comment le choix des unités lexicales et leur valeur sémantique varient en fonction du type de corpus et révèlent l'importance des type de discours sur les champs sémantiques, pour un thème donné.

2. Corpus : construction et présentation

Notre étude est basée sur un corpus composés de 2556 articles de presse recueillis dans trois sites :

a) institutionnel : www.gov.cn. Il s'agit d'un site officiel du gouvernement chinois ;

b) médiatique-institutionnel : www.people.com.cn. Il s'agit du site du Renmin Ribao, l'organe du Comité central du Parti communiste chinois. C'est le journal le plus prestigieux de la presse écrite nationale et le premier quotidien du pays ;

c) médiatique-informel : www.sohu.com. Ce site d'information générale est le plus grand portail web de la Chine.

Il est à noter qu'en Chine, les informations suspectées de mettre en danger la sécurité, la protection et la gestion du réseau d'information sont contrôlées par l'État dans tous les médias. Les articles publiés dans les sites médiatique-institutionnel et médiatique-informel sont tous contrôlés selon ce principe, mais suivant une permissivité variable.

Après une segmentation du corpus avec le module Python JIEBA, nous avons structuré le corpus en tenant compte des différents critères tels que le type de discours ou la date de publication du texte. Ces métadonnées nous permettent de multiplier les corpus (diachronique, discours, etc.).

1. Les termes spécifiques sont définis par leur spécificité calculée par l'algorithme implémenté (Lafon, 1980) dans Lexico5. Le calcul de la spécificité tient compte de l'écart observé entre la fréquence réelle d'une unité dans un sous-corpus donné rapportée à la fréquence théorique attendue dans le corpus total, les tailles des différents sous-corpus étant normalisées. Ce calcul nous informe donc de la probabilité qu'une unité apparaisse dans le sous-corpus. Il y a deux types de spécificité : positive ou négative. La spécificité positive indique le suremploi de l'unité, tandis que la spécificité négative montre le sous-emploi de l'unité dans le corpus par rapport à la valeur théorique attendue.

Type de corpus	Nombre d'articles	Nb d'occurrences	Nb de formes	Nb d'hapax	Type de métadonnées
Institutionnel (Gov)	1095	447183	28234	13196	9
Média-institutionnel (People)	835	506575	42824	21197	9
Média-informel (Sohu)	626	412772	34885	17077	9
Totalité des trois corpus	**2556**	**1366530**	**64654**	**31074**	**9**

TABLE 1 – Principales informations statistiques sur l'ensemble des sous-corpus

3. Méthodologie et résultats

3.1. Linguistique de corpus et textométrie

La linguistique de corpus permet d'observer les phénomènes linguistiques et les interpréter sur plusieurs aspects : la morphologie, la syntaxe ou la sémantique. La textométrie, comme sous-domaine de la linguistique de corpus, « propose des procédures de tris et de calculs statistiques pour l'étude d'un corpus de textes numérisés » (Pincemin 2011). Nous combinons ici la textométrie et l'analyse sémantique du corpus pour articuler les résultats statistiques et leur interprétation qualitative.

3.2. Présentation des premiers résultats

Gov (institutionnel)	People (média-institutionnel)	Sohu (média-informel)
影响 (impact, influence)	来源 (origine/source)	原因 (cause/raison)
F3650 f1796 S***	F367 f204 S14	F1327 f698 S***

TABLE 2 – Comparaison des trois unités lexicales spécifiques dans trois corpus

Parmi les éléments sémantiques les plus caractéristiques et les plus divergents des différents discours sur le wù maí, les causes apparaissent déterminantes. Nous observons que trois unités lexicales partageant en chinois des significations très proches, confinant à la synonymie, sont systématiquement utilisées dans les 3 corpus. En nous appuyant sur la concordance et les segments répétés de ces trois unités lexicales à l'aide de Lexico 5, nous présentons nos résultats d'analyse sémantique relative aux lexiques causes ou effets[2] du wù maí. En prenant comme exemple ces trois lexiques 影响 («influence ou impact»), 来源 («origines/sources»), 原因 («raison», «cause») surreprésentés dans leur partie correspondante, mais sous-représentées dans les autres, nous remarquons que les lexiques utilisés pour exprimer les causes du wù maí varient en fonction de leur type du corpus.

3.2.1. 影响 (influence ou impact) dans GOV institutionnel

Le corpus institutionnel (Gov) emploie le mot 影响 (« influence » ou « impact »). Il faut noter que le dernier mot offre non seulement des choix sur les catégories syntaxiques (il peut être un substantif

2. Seule 影响 («impact», «influence») apparu dans le corpus institutionnel évoque les effets et les causes du wù maí, tandis que 来源 (« origine ») et 原因 (« raison ») n'expriment que les causes.

ou un verbe selon le contexte d'utilisation), mais aussi en termes de catégories sémantiques : il explique soit les causes, soit les conséquences du phénomène visé. Dans le corpus Gov, la combinaison de 影响 (« influence ») et 雾霾 (« brouillard ») donne lieu à deux sens : les causes ou les effets du brouillard :

- 机动车 对 雾霾 的 影响(« l'influence des automobiles sur le wù maí »), i.e. les causes du brouillard
- 雾霾 天 对 呼吸系统 影响 最大(« les impacts les plus graves du wù maí portent sur le système respiratoire), i.e. les effets du brouillard PM2.5
- 对 人体 健康 和 大气环境 质量 的 影响 (les impacts du PM2.5 sur la santé et la qualité de l'air), i.e. les effets du brouillard

3.2.2. 来源 (origines/sources) dans PEOPLE média-institutionnel

L'unité lexicale sur-employée par le corpus médiatique-institutionnel (People) est 来源 (« origines/sources »). En tant que substantif, il sert à expliquer les causes du wù maí, par exemple :

- 雾霾的来源 (« les sources du wù maí »)
- 污染来源 (« les sources de la pollution »)
- PM2.5来源 (« les composants [sources] du PM2.5 »)

Dans ces exemples, nous notons qu'au lieu d'utiliser directement le mot wù maí, on emprunte l'abréviation technique désignant spécifiquement le problème de pollution, PM2.5 (pour *Particular Matter Ø 2.5 μm*). Soulignons que 来源 («origines/sources») est une unité plus littéraire et soutenue à l'écrit.

3.2.3. 原因 (raison/cause) dans SOHU média-informel

L'unité 原因 (« raison/cause ») utilisée dans le corpus média-informel (Sohu) est courante et relève de l'oral. Par exemple :

- 中国雾霾真正原因 (« les raisons originelles du wù maí en Chine »)
- 北京雾霾天气形成原因 (« les raisons du wù maí à Pékin »)
- 雾霾形成的原因 (« les raisons pour lesquelles le wù maí se produit »)

Dans ces trois exemples, les caractéristiques du corpus média-institutionnel se manifestent par l'ajout d'un complément circonstanciel de lieu à wù maí : 中国 (« Chine »), 北京 (« Pékin »). Cela permet de préciser l'endroit où se situe le problème de pollution.

4. Conclusion et perspective

À l'aide de l'approche textométrique, notre étude permet d'attester qu'à travers la concordance et les segments répétés des trois mots-pôles, qui sont sémantiquement proches, le choix du vocabulaire désignant un phénomène varie selon le type de discours du corpus. Les résultats nous donnent également la possibilité de voir *comment les différentes nominations de wù maí varient dans les 3 types de corpus et quel rôle joue l'ajout du complément circonstanciel de lieu.* Ces derniers constats nous permettront d'approfondir nos études avec pour objectif d'expliquer d'où viennent ces divergences. Nous envisagerons aussi de diversifier les thèmes analysés dans notre corpus (exemple : les enfants, les femmes, etc.) afin d'étudier les différentes manières d'appréhender le *wù maí* suivant les types de discours.

Références bibliographiques

Chetouani L. (2007). *« Les mots de la controverse sur le changement climatique »* , Le Télémaque 2007/1 n° 31 : 81-104.

Habert, B., Nazarenko, A. & Salem, A., 1997, *Les linguistiques de corpus.* Paris, Armand Colin - Masson.

Fu Huai-qing. 符淮青. 同义词研究中的几个问题 (2000). (Quelques problèmes dans les études des synonymes). 《中国语文》 Chinese Language.

Lafon P. (1980). *« Sur la variabilité de la fréquence des formes dans un corpus »*, Mots (n°1) : 127-165.

Lehmann, A., F. Martin-Berthe (1998), *«Introduction à la lexicologie. Sémantique et morphologie»*, Paris, Dunod, Coll.

LEXICO5, logiciel de textométrie disponible sur : http://www.lexi-co.com/L5Presentation.html

JIEBA 结巴中文分词. "Jieba" (Chinese for "to stutter") Chinese text segmentation : built to be the best Python Chinese word segmentation module. Logiciel disponible sur : https://github.com/fxsjy/jieba

Marie Veniard, Serge Fleury. *« Les manifestations textométriques de la saillance lexicale. Expérimentations et tentative de caractérisation ».* JADT - Journées internationales d'Analyses statistiques des Données Textuelles, 2016, Nice, France. Actes des Journées d'Analyses de Données Textuelles, Nice, juin 2016, 2016.

Pincemin, B. (2011) *« Sémantique interprétative et textométrie – Version abrégée »*, *Corpus*, 10, 259-269.

Pincemin, B., HEIDEN Serge (2008) – *« Qu'est-ce que la textométrie ? Présentation »*, Site du projet Textométrie, http://textometrie.ens-lyon.fr/spip.php?rubrique80

Rastier François (2005), *« Enjeux épistémologiques de la linguistique de corpus ».* In *La Linguistique de Corpus* (p.31-45). Presse universitaire de Renne.

Acquisition de la compétence de production écrite en FLE par des apprenants chinois : l'exemple de l'essai argumenté

Catherine David et Tatiana Aleksandrova
Université Grenoble Alpes
catherine.david@univ-grenoble-alpes.fr, tatiana.aleksandrova@univ-grenoble-alpes.fr

1. Introduction

L'acquisition de l'écrit en langue étrangère pose de nombreuses difficultés reconnues par les chercheurs en linguistique et en didactique. La maîtrise de la morphosyntaxe ne suffit pas à produire des textes cohérents de différents types et genres. Deux types de contraintes interfèrent dans ce processus : d'une part le processus psychologique d'organisation des idées, lié à la structure de la langue maternelle, et d'autre part la connaissance des règles du genre textuel propres à la culture et la langue cible.

2. Cadrage théorique

Selon le modèle psycholinguistique de production langagière proposé par Levelt (1989), la production d'une suite d'énoncés qui constitue un texte ou un discours, est le résultat d'un travail complexe de conceptualisation qui comporte plusieurs étapes et nécessite l'activation des connaissances culturelles, la prise en compte de la situation de communication, la planification des idées. Les chercheurs du domaine de l'acquisition des langues qui s'appuient régulièrement sur ce modèle, ont montré que les moyens linguistiques disponibles dans la langue du locuteur influencent la façon d'organiser les informations, notamment les mettre en ordre et les relier (Slobin, 1996, Lambert, 2006, von Stutterheim, 2003). Les apprenants d'une nouvelle langue doivent apprendre à la fois les moyens linguistiques et les processus d'organisation des informations pour construire différents types et genres de textes.

La rhétorique contrastive étudie les différentes structures textuelles relatives aux types et genres de textes en fonction des cultures et met en lumière la variabilité culturelle des pratiques d'écriture (Hidden, 2008, 2014). Ces études proposent des éclaircissements sur les difficultés propres aux apprenants de différentes langues maternelles (Bi, 2015), mais ne fournissent pas de dispositifs d'enseignement/apprentissage suffisamment étoffés. A son tour, la didactique des langues étrangères et du FLE est riche en manuels pédagogiques pour la production écrite mais ces derniers ne tiennent pas compte des aspects contrastifs.

Notre étude a pour objectif de faire le lien entre les processus psycholinguistiques mis en œuvre dans l'organisation des discours et la didactique du FLE à travers l'analyse contrastive des productions écrites. Cette visée didactique est justifiée par les besoins pour certains étudiants étrangers d'intégrer l'université française. Cette intégration est conditionnée par la réussite aux diplômes universitaires d'études de langue française (DUEF) ou aux diplômes d'études/aptitudes en langue française (DELF/DALF) qui exigent une capacité à exprimer son point de vue en français de façon structurée. Le texte argumentatif a donc retenu particulièrement notre attention. En effet, ce type de texte nécessite une structuration logique claire. Cependant elle peut être organisée de différentes manières comme en témoignent de nombreux modèles théoriques (Adam, 1990, Toulmin, 1993, Plantin, 2005). En tant qu'apprenant de notre LM nous intégrons ces modèles lors de la scolarisation. Mais en tant qu'apprenant d'une langue étrangère dont la culture est éloignée, les normes de l'écrit sont perçues comme des obstacles persistants.

3. Méthodologie

Notre choix s'est porté sur le public chinois de niveau avancé (B2-C1) qui est à la fois culturellement très éloigné du français et très présent dans les universités françaises. Nous envisageons par la suite d'étendre notre recherche à d'autres nationalités. Pour le moment, nous avons recueilli un corpus de 100 copies du DUEF/DELF d'apprenants chinois. Ce sont des essais argumentatifs d'environ 240 mots chacune. Afin de mieux comprendre leurs difficultés et les attentes des évaluateurs, nous avons envisagé de constituer des groupes de contrôles composés de locuteurs natifs chinois et français. Ces

derniers seront amenés à exécuter la même tâche de production écrite dans des conditions similaires. Le nombre de participants est estimé à 20 personnes par groupe. Le recueil de ces données est en cours. Nous disposons pour ces sujets d'informations socio-biographiques en leur proposant de remplir un questionnaire.

Le corpus sera numérisé et les productions en chinois seront glosées afin d'avoir une information sur la nature grammaticale de l'élément et sa traduction en français. Toutes les données seront analysées avec le même cadre d'analyse à savoir le modèle de la quaestio (Klein, von Stutterheim, 1993) issu de l'approche psycholinguistique à l'analyse du discours et les outils de la linguistique textuelle, à savoir des modèles qui schématisent le discours argumentatif. Quant au modèle de la quaestio, il a été largement utilisé pour l'analyse des productions orales de différents types. Ce modèle propose de voir toute production en tant que réponse à une question explicite ou implicite qui guide le locuteur ou le scripteur tout au long de sa production. En fonction de la quaestio, le discours peut être divisé en premier et second plan. Un découpage au niveau des énoncés peut également être réalisé entre les éléments du thème et du rhème. Cette approche fonctionnaliste nous permettra d'analyser les productions du point de vue de leur organisation informationnelle et sera complétée par une analyse textuelle.

4. Résultats

Les analyses contrastives des productions des groupes de contrôles doivent nous fournir des informations sur les caractéristiques des textes argumentatifs dans les deux langues. Ces différences nous serviront de référence pour mieux appréhender les spécificités des productions d'apprenants. Cette étape nous semble indispensable avant de pouvoir proposer des dispositifs pédagogiques.

Références bibliographiques

Adam, J.-M. (1992). *Les textes : types et prototypes.* Paris : Nathan.

Bi, X. (2015). *Rhétorique de la dissertation. Etude contrastive des conventions d'écriture académique en français et en chinois,* Thèse de Doctorat, Université Paris 3.

Hidden, M.-O. (2008). Apprendre à rédiger un texte en français L2. *Acquisition et interaction en langue étrangère, 27,* 109-122.

Hidden, M.-O. (2014). *Pratiques d'écritures, apprendre à rédiger en langue étrangère,* Paris : Hachette.

Klein, W. & Stutterheim, von Ch. (1991). Text structure and referential movement. *Sprache und Pragmatik, 22,* 1-32.

Lambert, M. (2006). Pourquoi les apprenants adultes avancés ne parviennent-ils pas à atteindre la compétence des locuteurs natifs ? In Engwall, G. (éd.) *Construction, acquisition et communication : études linguistiques des discours contemporains.* Acta Universitatis Stockholmiensis, Romanica Stickholmiensia, 23, 151-171.

Levelt, W., J., M. (1989). *Speaking : From Intention to Articulation.* Cambridge : MIT Press.

Plantin, C. (2005). *L'argumentation,* Paris : PUF.

Slobin, D. I. (1996). From "thought and language" to "thinking for speaking". In Gumperz, J.J. & Levinson, S. C. (éds.) *Rethinking linguistic relativity.* Cambridge : Cambridge University Press, 70-96.

von Stutterheim, Ch. (2003). Linguistic structure and information organisation : the case of very advanced learners. In Foster-Cohen, S. & Pekarek Doehler, S. (éds.) *EUROSLA yearbook.* Amsterdam/Philadelphia : John Benjamins, 183-206.

Toulmin, S. (1993). *Les usages de l'argumentation,* trad. P. de Brabanter, Paris : PUF.

Corpus en classe de FLE : difficultés et propositions pédagogiques. L'exemple avec les prépositions

Thi Thu Hoai Tran [1] et Rui Yan [2]
[1]Grammatica, Université d'Artois
[2]Lidilem, Université Grenoble Alpes

Notre travail de recherche se situe dans la lignée des travaux de Chambers (2010), Di Vito (2013), Boulton et Tyne (2014), en mettant l'accent sur l'introduction du corpus en classe de langue. En fait, il a été démontré que les corpus se montrent efficaces par apport à d'autres formes de pratiques dans différents contextes d'enseignement (Cobb et Boulton, 2015). Comme nous souhaitons mettre à disposition des étudiants allophones une aide à la rédaction scientifique, un travail sur les articles scientifiques s'avère nécessaire pour les aider à se familiariser à un nouveau genre d'écrit et aux nouvelles structures. A notre avis, l'introduction du corpus en classe de langue s'effectue à deux niveaux, le premier niveau concerne l'enseignant qui doit prendre conscience de l'utilité d'un apprentissage sur corpus. Cependant, Cavalla et Loiseau (2014) ont remarqué avec justesse la réticence des enseignants face à l'utilisation de cet outil. Le deuxième niveau renvoie aux apprenants, une surcharge cognitive ainsi que le développement de l'autonomie doivent être pris en considération.

Ce travail a pour objectif principal d'encourager les enseignants à utiliser le corpus en classe de langue et de souligner l'apport des analyses linguistiques pour des applications didactiques. Nous tentons de montrer que le corpus permet, d'une part, à l'enseignant d'alimenter le cours grâce aux documents authentiques et, d'une autre part, aux apprenants de mener une réflexion métalinguistique et de s'approprier les associations de mots. En effet, de nombreuses études ont montré que les corpus sont utiles notamment pour le public de niveau avancé. Pourtant, dans ce travail, nous nous interrogeons sur la manière d'introduire le corpus pour un public de bas niveau.

Dans le cadre de notre formation du Diplôme universitaire Français Langue Etrangère pour la Préparation aux Etudes Supérieures, nous pouvons travailler avec un public varié qui vient de différents pays et dispose d'un bagage linguistique très hétérogène. Ces étudiants souhaitent s'intégrer à différentes filières (économie, finance, sciences du langage, génie civil, etc.). Ils suivent un an de préparation linguistique dans l'intention d'obtenir le DELF B2, condition pour poursuivre leurs études à l'Université. En plus des cours sur les techniques universitaires, ces étudiants ont également des cours de soutien linguistique afin de combler certaines lacunes en langue française. Nous avons constaté que l'utilisation des prépositions fait partie des difficultés les plus récurrentes chez les étudiants. C'est pourquoi, cette communication s'intéresse essentiellement à l'enseignement des prépositions à ce public tout en nous basant sur le corpus.

Dans une recherche antérieure, nous avons travaillé sur le corpus composé d'articles scientifiques en sciences humaines et sociales, environ 5 millions de mots (Tran, 2014, Hatier, 2016). Ce corpus a été étiqueté morpho-syntaxiquement et annoté semi-automatiquement grâce aux techniques du Traitement Automatique des Langues. Les études portaient sur le lexique scientifique transdisciplinaire, un lexique de genre qui transcende toutes les disciplines (Tutin, 2007). L'analyse des constructions verbales nous permet de relever les propriétés syntaxiques et sémantiques du lexique scientifique verbal et de circonscrire le réseau sémantique établi par les prépositions. Il s'en résulte que certaines prépositions associées à des constructions verbales spécifiques participent à la construction des sens. Il nous semble donc important de travailler à la fois les constructions verbales et leur interprétation sémantique.

Le tableau ci-dessus récapitule quelques utilisations des prépositions les plus fréquentes. Par exemple, la préposition *à* est souvent utilisée dans des constructions qui indiquent une relation d'appartenance avec un autre élément. La préposition *de* est utilisée pour indiquer la source, la préposition *sur* pour désigner le fondement. Il est donc indispensable de sensibiliser les étudiants à ce phénomène linguistique. L'apprentissage se fait dans une approche inductive où les apprenants vont être amenés à découvrir eux-mêmes ces différents champs sémantiques.

Dans le cadre de notre expérimentation avec les étudiants allophones qui disposent d'un bagage linguistique très différent, nous avons constaté que les exemples venant des articles scientifiques restent difficiles par rapport au niveau des étudiants. Comme nous envisageons de faire travailler les étudiants

Prépositions	Sens	Verbes
A	[Appartenance] : relation renvoyant pour un individu au fait d'appartenir à un milieu, une collectivité.	*Appartenir, prendre part*
	Processus humain / attribution : processus par lequel on donne une propriété à un élément.	*Accorder, apporter, attribuer*
DE	[Source] : relation mettant en lien l'origine, la raison de quelque chose et le résultat	*Résulter, découler, hériter, provenir, ressortir*
SUR	[Fondement] : action scientifique qui vise à prendre quelque chose pour base.	*Se fonder, se baser, s'appuyer, reposer*
DANS	[Inclusion] : l'état de comprendre, de réunir en soi par nature des éléments ou des caractères qualitatifs ou quantitatifs.	*Inclure, insérer, introduire*

essentiellement sur le sens des prépositions, plus précisément, sur l'association des verbes et les prépositions afin de faire ressortir les régularités des prépositions, il est donc important de travailler sur les exemples facilement accessibles. Afin de faciliter l'accès au corpus, nous envisageons de travailler sur le corpus de vulgarisation scientifique qui représente pour nous une transition avant d'entrer au corpus d'écrits scientifiques, surtout pour un public de bas niveau. Ce corpus composés de 125 articles, issus de la revue Sciences Humaines [1] permettent de faire travailler les apprenants sur le lexique scientifique transdisciplinaire sans se trouver confrontés à la question de terminologie.

Références bibliographiques

Boulton, Alex, et Henry, Tyne (éds). *Des documents authentiques aux corpus. Démarches pour l'apprentissage des langues.* Paris : Didier, 2014.

Bruley-Meszaros, Cécile. Quel enseignement des verbes en didactique du français langue étrangère ?. *Synergies France, 6,* 51-59, 2010.

Cavalla, Cristelle et Mathieu Loiseau. « Scientext comme corpus pour l'enseignement ». In A. Tutin et F. Grossmann (éds),*L'écrit scientifique : du lexique au discours. Autour de scientext,* PUG., 163-80. Rennes, 2013.

Chambers, Angela. « L'apprentissage de l'écriture en langue seconde à l'aide d'un corpus spécialisé ». *Revue française de linguistique appliquée,* XV, 9-20, 2010.

Hatier, Sylvain. « Extraction et catégorisation de lexiques scientifiques transdisciplinaires d'articles scientifiques de sciences humaines en vue de l'indexation automatique ». Thèse de doctorat, Université de Grenoble Alpes, 2016.

Homma, Yukiyo. Analyse critique et révision de quelques points de vue théoriques sur l'alternance entre A et DANS en vue d'une problématique de l'enseignement des prépositions françaises en FLE. In : *Colloque international Recherches en acquisition et en didactique des langues étrangères et secondes,* 2006.

Riegel, Martin, Jean-Christophe, Pellat, et René, Rioul. *Grammaire méthodique du français.* 7e édition revue et augmentée. Linguistique nouvelle. Paris : Presses universitaires de France, 2009.

Sonia, Di Vito, « L'utilisation des corpus dans l'analyse linguistique et dans l'apprentissage du FLE », *Linx* [En ligne], 68-69, 2013.

Tom, Cobb et Alex, Boulton. « Classroom applications of corpus analysis ». In D. Biber & R. Reppen (éds), *Cambridge Handbook of English Corpus Linguistics.* Cambridge : Cambridge University Press, 478-497, 2015.

Tran, Thi Thu Hoai. « Développement d'une aide à l'écrit scientifique. Description de la phraséologie scientifique et réflexion didactique pour l'enseignement à des étudiants non natifs. » Thèse de doctorat en Sciences du langage Spécialité Français Langue Etrangère, Université Grenoble Alpes, 2014.

Tutin, Agnès. *Lexique et écrits scientifiques.* Vol. XII-2. Revue Française de Linguistique Appliquée, 2007.

1. Téléchargeable depuis le site ORTOLANG (Open Resources and TOols for LANGuage) : https://hdl.handle.net/11403/scienceshumaines/v1

La structure V+*difficulté(s)* et ses emplois dans le discours scientifique des orthophonistes

Frédérique Brin-Henry [1 2] et Marie Laurence Knittel [2]
[1]ATILF- UMR 7118 / Université de Lorraine
[2]Centre Hospitalier de Bar-le-Duc
frederique.henry@atilf.fr, marie-laurence.knittel@univ-lorraine.fr

1. Introduction

La terminologie orthophonique contribue à la communication efficace entre les professionnels de santé, permet la mise en mots des difficultés du patient, et peut être considérée comme un témoignage des représentations de la société sur la nature et l'impact du handicap que constituent les troubles du langage. Depuis 2014, nous travaillons à l'exploration et l'analyse syntactico-sémantique des occurrences de certains termes dans les textes. Ainsi est apparue la fréquence remarquable du nom *difficulté(s)*, dont nous avons examiné les contextes droits dans un corpus de comptes rendus de bilans orthophoniques (Brin-Henry & Knittel 2015, 2016). Nous pensons que le nom *difficulté(s)* acquiert dans la littérature orthophonique un statut terminologique (L'Homme, 2004), et nous interrogeons la spécificité de l'usage de tels noms, qui endossent indéniablement un rôle particulier dans les routines du discours orthophonique écrit (Née *et al*, 2012). Dans le présent travail nous étendons notre examen à l'étude des verbes introducteurs (voir L'Homme 2012) et vers un nouveau corpus d'écrits en orthophonie.

2. Corpus et Méthodologie

2.1. Présentation du corpus

Notre corpus [1] regroupe 850 articles de la revue *Rééducation Orthophonique*, rédigés à 59% par des orthophonistes (écrivant seul ou avec d'autres professionnels) et compte environ 3 millions de mots. Ce corpus a été intégré à la plateforme de textométrie TXM [2] (Heiden *et al.* 2010). Le Tableau 1 présente quelques données descriptives montrant les variations formelles observées.

	Item	**Nombre**	**Exemple**
Documents	Articles	850	
	Numéros de revue	68	
	Période(18 années)	1997 - 2014	
Contenu	Tokens	4397847	
	Types de tokens	103921	*difficulté* vs. *difficulté**s*** vs. ***D**ifficulté* vs. ***D**ifficulté**s***
	Types de formes lemmatisées	72613	*difficulté* vs. *difficile*
	Catégories grammaticales (TreeTagger)	33	

TABLE 1 – **Données descriptives du corpus d'articles de la revue Rééducation Orthophonique**

2.2. Méthodologie de la recherche sur corpus : outils et procédure

L'extraction des occurrences de *difficulté(s)* s'est faite selon plusieurs modalités successives. Dans un sous-corpus d'articles rédigés par au moins un orthophoniste, nous avons repéré les verbes situés

1. Consitué dans le cadre du projet ORTHO CORPUS 2015-2017, co-financé par le laboratoire ATILF, la Région Lorraine, la Fédération Nationale des Orthophonistes et le concours de l'éditeur Ortho-Edition

2. La plateforme TXM, en combinant des techniques puissantes et originales, en synergie avec les technologies de corpus et de statistique actuelles (Unicode, XML, TEI, TAL, CQP et R), permet l'analyse de grands corpus de textes au moyen de composants modulaires et open-source (http://textometrie.ens-lyon.fr/).

Verbe en contexte gauche	Fréquence
ont des difficultés	23
présentent des difficultés	22
présentant des difficultés	21
ayant des difficultés	19
a des difficultés	10
est en difficulté	8
sont en difficulté	7
éprouve des difficultés	6
est une difficulté	5
met en évidence des difficultés	5

TABLE 2 – **Principaux verbes situés entre 0 et 5 mots à gauche du lemme difficulté**

entre 0 et 5 mots à gauche du lemme *difficulté*[3]. Nous avons recensé 1748 occurrences correspondant à 1516 formes fléchies (voir le Tableau 2 pour quelques exemples), parmi lesquelles nous avons repéré des structures dans lesquelles *difficulté(s)* apparait comme objet direct ou indirect d'un verbe fléchi, ou dans une construction existentielle (*il y a, il s'agit, il existe*). Notre choix d'étudier les verbes à gauche de *difficulté(s)* se justifie par le fait que nous souhaitions porter notre attention sur les verbes introducteurs de ce nom, notamment les verbes-supports (voir sections 3.2. et 4), majoritairement situés devant le nom. Les contextes droits de *difficulté(s)*, y compris les subordonnées infinitives introduites par à (*difficultés à comprendre*), ayant fait l'objet de travaux antérieurs (Brin-Henry & Knittel, 2016), ils n'ont pas été traités ici.

3. Données linguistiques

3.1. Le nom difficulté(s)

Le nom d*ifficulté(s)* est un nom désadjectival construit sur l'adjectif *difficile*. Comme de nombreux noms de cette classe, *difficulté(s)* possède deux emplois. Employé comme massif, il dénote la propriété d'être difficile (Beauseroy 2009), qui s'applique, comme avec l'adjectif, à la réalisation d'une action, explicitée par un verbe ou un nom (*lire*/*lecture*).

1. a. {la lecture / lire} est difficile = b. la difficulté de {la lecture / lire}

Employé comme comptable, en particulier au pluriel, *difficulté* renvoie plus spécifiquement à un argument de l'adjectif caractérisé par la propriété dénotée par l'adjectif lui-même (voir Knittel 2015).

2. a. Les difficultés {du livre / de la lecture} = b. Ce qui est difficile dans {le livre / la lecture}

Difficulté(s) et *difficile* peuvent présenter un argument supplémentaire introduit par *pour*, correspondant à l'entité qui ressent la difficulté, et interprété comme un *Experiencer*

3. a. La lecture est difficile pour cet élève b. {la/les} difficulté(s) de la lecture pour cet élève

3.2. La notion de verbe-support

Nous adoptons ici une conception large de la notion de verbe-support, que nous caractérisons par deux propriétés majeures (pour une conception plus restreinte, voir Vivès 1993). D'une part, il s'agit d'un verbe faiblement porteur de sens, au contraire d'un verbe lexical 'plein' ; d'autre part, cette classe de verbes se combine avec un nom prédicatif (Giry-Schneider, 1978 ; Gross, 1981 ; Danlos, 2009 ; Vivès

3. Verbes situés entre 0 et 5 mots à gauche du lemme difficulté, dans un sous-corpus d'articles écrits par des orthophonistes uniquement : Index de <[frpos='V.*'][]0,5[frlemma="difficult.*"]> avec la propriété [word] dans le corpus ORTHOCORPUS_auteurorthophoniste

1984) en position objet. Ce nom prédicatif contribue fortement à la sélection et à l'interprétation de l'argument externe. Dans le cas de *difficulté(s)*, le choix du verbe-support influence l'interprétation du nom. Ainsi :

- Le verbe *avoir* nous semble sémantiquement neutre.
- Les 'verbes de ressenti' *éprouver* et *ressentir* permettent de présenter le sujet comme un Experiencer, et caractérisent *difficulté(s)* comme un nom de sentiment (Anscombre 1995).
- Les 'verbes de manifestation' (*montrer, manifester, présenter, faire preuve de et témoigner de*), classent *difficulté(s)* parmi les attitudes, perceptibles de l'extérieur (Anscombre, 1995 ; Goossens, 2011).
- La structure *être en*, typique des prédicats d'états (Van de Velde, 1995 ; Flaux & Van de Velde 2000) présente *difficulté(s)* comme un état, localisant l'individu auquel réfère le sujet.

4. La distribution des verbes devant *difficulté(s)*

Les observations qui suivent résultent de l'examen des 1748 occurrences des verbes sous forme fléchie en contexte gauche de *difficulté(s)* dans notre corpus, et de leur fréquence. Nous étudions d'une part la distribution des verbes-supports ci-dessus afin de voir si *difficulté(s)* est plutôt présenté comme un nom de sentiment, d'attitude ou d'état. Nous examinons d'autre part les verbes (supports ou non) les plus fréquents dans nos structures, supposant qu'ils révèlent ce que disent les orthophonistes de(s) *difficulté(s)* dans leurs écrits scientifiques.

4.1. Distribution des verbes-supports

La distribution des verbes-supports répertoriés est présentée dans le Tableau 3. De façon attendue, le verbe le plus fréquent est *avoir*, sémantiquement neutre, et transmettant une vision exogène de la difficulté. Les autres classes de verbes sont très minoritaires (moins de 15% pour chacune). Quelques tendances peuvent être relevées :

Pour véhiculer le point de vue du sujet / Experiencer, les scripteurs utilisent majoritairement *éprouver* plutôt que *ressentir* (33 occurrences vs 1 occ.). Pour attester de l'existence des difficultés, les verbes majoritaires sont *présenter* (17 occ.) et *montrer* (11 occ.). Enfin, *difficulté(s)* n'est que peu présenté comme un état (23 occ. sur le total de 242).

Verbes	**Nombre d'occurences**	**Pourcentage**	**Types de verbes**
éprouver	33	13,7	**V de ressenti** : 14,01%
ressentir	1	0,4	
montrer	11	4,6	**V de manifestation** : 12,7%
manifester	3	0,8	
présenter	17	7,03	
faire preuve de	0	0	
témoigner de	0	0	
être en	154	63,7	**Verbe introducteur d'état** : 9,5%
avoir	154	63,7	63,7%
Total	242	100%	

TABLE 3 – **Fréquence des verbes-supports devant *difficulté(s)***

4.2. Distribution selon la fréquence

Le Tableau 4 présente les verbes dont le nombre d'occurrences est supérieur ou égal à 10, toutes flexions confondues. On note la présence de verbes-supports et d'autres verbes davantage porteurs de sens. Nous avons réparti ces verbes en 4 groupes selon ce à quoi ils renvoient (cf. Levin 1993) : mise en évidence, analyse, cause et évolution, la fréquence devant être relativisée par la prise en compte de (quasi)synonymes.

Ceci permet de proposer une nouvelle classification reposant sur la fréquence par type d'information véhiculé. Dans le Tableau 5, on constate ainsi que ce sont la mise en évidence des difficultés (66 verbes et locutions) et la détermination de leurs causes (48 verbes et locutions) qui constituent les deux orientations majeures de l'emploi de ce mot-clé. Ces données permettent d'évoquer des routines discursives (Veniard 2008) dans ce corpus d'écrits scientifiques élaborés par les orthophonistes.

Verbes	**Nombre d'occurences**	**Verbes**	**Nombre d'occurences**
avoir	154	connaître	12
éprouver	33	décrire	12
être en	23	monter	11
entraîner	22	pallier	11
présenter	17	compenser	10
expliquer	15	être (une difficulté)	10
mettre en évidence	14	noter	10
comprendre	12		

TABLE 4 – **Verbes de fréquence supérieure à 10 devant *difficulté(s)***

	Verbes de fréquence > 10	**autres verbes (fréquence)**	**Total**
Mise en évidence	noter (10), mettre en évidence (16), décrire (12)	constater (5), observer (3), détecter (2), identifier (2), il y a (2), mettre à jour (2), repérer (2), déceler (1), démontrer (1), entrevoir (1), faire état (1) indiquer (1), isoler (1), mentionner (1), mettre en avant (1), pointer (1), signaler (1)	66
Analyse	expliquer (15), comprendre (12)	analyser (5), apporter une explication (1), interroger (1)	34
Cause	entraîner (22)	engendrer (6), induire (3), poser (3), mettre en (2), conduire à (2), générer (2), attribuer la difficulté à (1), donner lieu à (1), faire naître (1), faire apparaître (1), provoquer (1)	45
Évolution	pallier (11), compenser (10)	augmenter (3), dépasser (3), faire disparaître (2), résoudre (2), contrer (1), amender (1), prévenir (1), traiter (1), vaincre (1)	36

TABLE 5 – **classes de verbes les plus fréquentes devant *difficulté(s)***

5. Conclusion

Notre étude montre que dans notre corpus, *difficulté(s)* est prioritairement introduit par *avoir*, verbe-support qui en signale simplement la présence. Les autres verbes les plus fréquents renvoient à la mise en évidence, l'analyse, les causes et l'évolution de la / des difficulté(s), et font émerger certains motifs des écrits scientifiques des orthophonistes.

Références bibliographiques

Anscombre, J.C. (1995). Morphologie et représentation événementielle : le cas des noms de sentiment et d'attitude. *Langue Française 105*, 40-54.

Beauseroy D. (2009). *Syntaxe et sémantique des noms abstraits statifs. Des propriétés verbales et adjectivales aux propriétés nominales.* Thèse de doctorat, Nancy-Université.

Brin-Henry, F. (2014). Using corpus-based analyses in specialised paramedical French. *Revue Française de Linguistique Appliquée : Langues de spécialité : problèmes et méthodes 19-1*, 103-15.

Brin-Henry, F., Knittel, M.L. (2015). L'usage des termes *difficulté(s)* et *trouble(s)* dans un corpus de comptes rendus de bilans orthophoniques. Communication présentée lors du colloque Cures de langage(s), Arras : 10-11 Décembre 2015.

Brin-Henry F., Knittel, M.L. (2016). Etude lexicosémantique du nom *difficulté(s)* dans les comptes rendus de bilan orthophonique : apports structuraux et conceptuels. *LIDIL 53*, 19-41.

Danlos, L. (2009). Extension de la notion de verbe support. *Actes du Colloque International Supports et prédicats non verbaux dans les langues du monde*. Paris. 28-33.

Flaux, N., Van de Velde, D. (2000). *Les noms en français : esquisse de classement*. Paris : Ophrys.

Gross, M. (1981). Les bases empiriques de la notion de prédicat sémantique. *Langages 63*, 7-52.

Giry-Schneider, J. (1978). *Les nominalisations en français : l'opérateur faire dans le lexique*. Genève : Droz.

Goossens V. (2011). *Propositions pour le traitement de la polysémie régulière des noms d'affect*. Thèse de doctorat, Université de Grenoble.

Heiden, S., Magué, J-P., Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In I. C. Sergio Bolasco (Ed.), *Proc. of 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*). Rome : Edizioni Universitarie di Lettere Economia Diritto, Vol. 2, 1021-1032.

Knittel, M.L. (2015). Adjectifs et noms désadjectivaux comptables : quelles relations ? *Verbum* 37-1, 91-114.

Levin, B. (1993). *English verb classes and alternations*. Chicago : University of Chicago Press.

L'Homme, M. C. (2004). *La terminologie : principes et techniques*. Montréal : Presses Universitaires de Montréal.

L'Homme, M. C. (2012). Le verbe terminologique : un portrait de travaux récents. In Neveu, F. et al. (éd). *Actes du 3e Congrès mondial de linguistique français*e, Lyon.

Nee, E., Sitri, F., Veniard, M. (2014). Pour une approche des routines discursives dans les écrits professionnels. In Neveu, F et al. (ed) *Actes du 4º Congrès Mondial de Linguistique Française*, Berlin.

Van de Velde, D. (1995). *Le spectre nominal. Des noms de matières aux noms d'abstractions*. Louvain / Paris : Peeters.

Veniard, Marie. (2008). Ecrire "ce qui ne va pas" dans le champ de l'enfance en danger : les mots *problème(s)* et *difficulté(s)*. *Carnets du Cediscor 10*, 57-77.

Vivès, R. (1984). *Perdre*, extension aspectuelle du verbe support a*voir*. *Revue québécoise de linguistique* 13-2, 13-57.

Vivès, R. (1993). La prédication nominale et l'analyse par verbes supports. *L'information grammaticale 59*, 8-15.

Les différents discours du domaine « Climat et énergie » en espagnol, à l'épreuve de la linguistique de corpus

Thierry Nallet et Sandrine Rol-Arandjelovic
Université Grenoble Alpes
prenom.nom@univ-grenoble-alpes.fr

Dans le prolongement d'une première recherche sur les discours de la presse espagnole et latino-américaine portant sur l'Accord de Paris, nous avons mis en œuvre cette année, avec l'aide d'étudiants stagiaires dans le cadre d'un financement du consortium « CORpus, Langues et Interactions» (CORLI), la constitution d'un corpus monolingue autour de la COP21, la 21e conférence des parties qui a eu lieu à Paris en décembre 2015.

Ce corpus en espagnol se compose de trois sous-corpus homogènes caractérisés par un genre textuel différent. Le premier rassemble plus de 800 articles de presse (555 000 mots). Le second regroupe 43 rapports produits par des ONG nationales et internationales (plus d'un million de mots). Le dernier, composé de 27 textes et documents officiels rédigés par des experts de l'ONU, compte plus de deux millions de mots. Au total, le corpus comprend 3 656 000 mots, ce qui permet une large exploration du domaine.

Dans le cadre de cette recherche sur la langue liée à la spécialité « Climat et énergie », il nous a semblé intéressant de remonter au point de départ de la COP21 : la tenue en 2014 de la COP20 à Lima tout en prolongeant cette étude jusqu'à aujourd'hui, afin de pouvoir appréhender les conséquences qu'a pu avoir l'Accord de Paris et d'inclure la COP 22 qui s'est déroulée à Marrakech en décembre 2016.

Les types de discours portant sur la COP21 forment au niveau « macro », celui du corpus total, un ensemble cohérent et représentatif du changement climatique et des politiques énergétiques. Cependant, la méthodologie employée de type « corpus-driven » permettra la libre exploration « micro » de traits saillants liés à un genre textuel déterminé grâce à un traitement indépendant de chaque sous-corpus. Cette exploitation s'appuiera notamment sur l'outil de textométrie TXM, en complément de l'extraction terminologique effectuée à l'aide de TermoStat. Dans les limites de notre poster, nous mettrons en évidence quelques traits notables se rapportant à la langue de spécialité, dans une optique tant quantitative que qualitative. L'analyse des différents types de discours nous permettra de voir des phénomènes de « répétition » *vs* « variation ». Nous nous intéresserons tout particulièrement à la néologie et à la déterminologisation, c'est-à-dire le passage dans la langue commune de termes spécialisés. L'hétérogénéité générique abordée mettra en évidence *in fine* dans quelles mesures il existe une communauté de discours autour du domaine.

Références bibliographiques

ADAM, J-M. (1999). *Linguistique textuelle. Des genres de discours aux textes*, Paris : Nathan « Université ».

BÉJOINT, H., THOIRON, Ph. (dir.) 1999). *Le sens en terminologie*. Lyon : Presses universitaires de Lyon,.

BORDET, G (2015). Culture de recherche, culture professionnelle : retrouver des continuités par la formation à la traduction spécialisée, in *Carton, Narcy-Combes, Narcy-Combes, Toffoli (dir.), Cultures de recherche en linguistique appliquée*. Paris : Riveneuve éditions, p. 145-165.

BOULTON, A., TYNE, H. (2014). *Des documents authentiques aux corpus : démarches pour l'apprentissage des langues*. Paris : Didier, « Langues et didactique ».

FRÉROT, C. (janv-jun 2016). Corpora and corpus technology for translation purposes in professional and academic environments. Major achievements and new perspectives, *Cadernos de Tradução, Florianópolis, v. 36, nº especial* 1, p. 36-61.

GONZÁLEZ REY, M. I. (dir.) (2014). O*utils et méthodes d'apprentissage en phraséodidactique*. Bruxelles : EME, « Proximités - didactique ».

KÜBLER, N., VOLANSCHI, A. (2012). Semantic prosody and specialised translation, or how a lexico-grammatical theory of language can help with specialised translation , in *Boulton A., Carter-Thomas S. & E. Rowley-Jolivet (dir.), Corpus-informed Research and Learning in ESP : Issues and Applications*. Amsterdam/Philadelphie : John Benjamins, p. 105-135.

LOOCK, R. (2016). *La traductologie de corpus.* Villeneuve d'Ascq : Presses Universitaires du Septentrion.

MAINGUENEAU, D. (1991). *L'analyse du discours.* Paris : Hachette.

PETIT, M.(2006). Les descripteurs du cadre : quelle conception de la langue de spécialité ? , *in Haramboure et alii (dir), Travaux des journées 2006 de l'EA 2025.* Bordeaux : Université Victor Segalen Bordeaux 2, p. 14-29.

RASTIER, F. (juin 2004). Enjeux épistémologiques de la linguistique de corpus, *Revue Texto !, Rubrique Dits et inédits*, [en ligne], Disponible sur : http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html (Consultée le 10 décembre 2016).

TEUBERT, W. (2009). La linguistique de corpus : une alternative [version abrégée], *Semen* [en ligne], *27 |2009*, mis en ligne le 01 avril 2009, consulté le 27 juin 2016. URL:http://semen.revues.org/8914

Les pronoms *je* et *nous* dans l'*Encyclopédie* et dans *Wikipédia* : propos d'une comparaison

Tobias von Waldkirch
Université de Neuchâtel
tobias.vonwaldkirch@unine.ch

1. Introduction

La question du rapport entre le locuteur et son discours suscite des réflexions et des débats depuis longtemps. Ainsi, le 29 décembre 1757, Voltaire écrivit à d'Alembert au sujet de l'écriture dans l'Encyclopédie [1] :

> [...] chacun veut étendre ses articles. On oublie, comme dit Pascal, qu'on est ligne, et on se fait centre. On veut occuper une grande niche dans votre Panthéon : on ose dire *je* et *moi* dans votre dictionnaire. Ah ! Que je suis fâche de voir tant de strass avec vos beaux diamans !

La représentation du savoir et la place du locuteur dans son discours diffèrent d'une époque à l'autre. Pour étudier cela, nous avons choisi deux encyclopédies qui représentent deux états différents de la langue française : l'*Encyclopédie* de Diderot et d'Alembert pour le XVIIIe et *Wikipédia* pour le XXIe siècle. Ces deux œuvres ont été choisies en raison de leur caractère représentatif pour le discours encyclopédique de leur époque respective. L'*Encyclopédie* était innovatrice dans la mesure où elle suppléait aux préceptes encyclopédiques tributaires de la religion chrétienne par un fil conducteur qu'étaient la raison et la science. En outre, c'est la première encyclopédie à être rédigée par un comité d'auteurs et non plus par un seul érudit. De même, *Wikipédia* change radicalement le discours encyclopédique de nos jours puisque la distinction entre auteur et lecteur est rendue caduque grâce à la possibilité d'intervenir directement sur le texte. Ce qui est commun à ces deux œuvres, c'est d'avoir facilité l'accès au savoir et fait évoluer les rapports entre l'homme et la connaissance à leur époque respective.

Notre communication portera sur la position du locuteur *je* et *nous* dans ces deux corpus. Alors que dans l'*Encyclopédie* de Diderot et d'Alembert, on a l'impression que le locuteur se manifeste régulièrement à la première personne du singulier ou du pluriel, on ne pourrait guère s'imaginer qu'un locuteur dise *je* de nos jours dans une encyclopédie telle que *Wikipédia*. Notre communication proposera un regard sur cette problématique en soumettant les occurrences de je et nous dans l'*Encyclopédie* (23 940 181 mots) et dans *Wikipédia* (251 981 725 mots [2]) à des analyses quantitative et qualitative. La comparaison entre les deux cherchera à dégager les possibles sens et fonctionnements discursifs que ces pronoms peuvent avoir dans les deux corpus. Ensuite, nous tenterons de mettre en rapport les résultats de l'analyse avec le changement des courants épistémologiques entre le XVIII^{e} et le XXI^{e} siècles.

2. Corpus et méthodologie de l'analyse quantitative

Les recherches sont effectuées sur la plateforme BTLC qui permet l'accès à plusieurs corpus ainsi que l'analyse statistique des données recueillies [3]. Dans un premier temps, il sera question de présenter les différences statistiques relevées par le recensement des deux pronoms [4] :

Les champs F1 et F2 retiennent les fréquences absolues pour les items cherchés dans les deux corpus, dont le nombre de mots de chacun est indiqué par N1 et N2. Les fréquences relatives, RF1 et RF2, rendent possible la comparaison entre les deux corpus. Il s'avère que la présence de *je* et de *nous*

1. Cité dans PERRONNEAU, V. H et CÉRIOUX, A. (1821) (éds). *Œuvres complètes de Voltaire. Tome quarante-troisième. Correspondance avec M. d'Alembert.* Paris : Plassan, p. 54. Nous remercions Prof. Muriel Brot de nous avoir indiqué cette lettre.

2. Il s'agit d'un sous-corpus constitué par un article sur onze.

3. Dirigée par P. Blumenthal (Cologne), développée par S. Diwersy (Montpellier 3).

4. Tous les chiffres sont arrondis au centième près. Pour la commodité de la lecture, les valeurs de RF1 et RF2 ont été multipliées par mille.

	DDA=1	WIKI=2								
	F1	F2	N1	N2	E1	RF1	E2	RF2	LR	LL
Je	21 670	6 777	23 940 181	251 981 725	2 468,1919	9,05‰	25 978 ,82	0,026‰	5,07	75 940,65
Nous	45 802	8 436	23 940 181	251 981 725	4 705,92	19,13‰	49 532,08	0,033‰	5,84	178'580,08

Table 1 –

est nettement supérieure dans l'*Encyclopédie* : si elle s'y élève pour *je* à 9,05‰et pour nous à 19,13‰, les mêmes pronoms n'ont qu'une fréquence relative de 0,026‰et 0,033‰dans *Wikipédia*. Les valeurs E1 et E2 permettent en outre la comparaison des fréquences estimées : le chiffre E1 pour le pronom *nous* indique donc que le premier corpus (DDA) recenserait 4 705,92 occurrences (au lieu des 45 802 effectivement trouvées), si la présence de cet item y était à la même échelle que dans le deuxième corpus (Wiki). En revanche, si le taux d'occurrence de *nous* dans la *Wikipédia* était proportionné au taux de l'*Encyclopédie*, on y trouverait 49 532,08 occurrences (au lieu des 8 436 effectivement trouvées). S'élevant à 5,07 et 5,84, les valeurs des calculs LR dépassent même le seuil qui permet la conclusion qu'au sein de l'*Encyclopédie*, il y a 512 fois plus d'occurrences des pronoms *je* et *nous* comparé à *Wikipédia* (c'est-à-dire $5{,}07^2$, respectivement $5{,}84^2$). Finalement, les valeurs log-likelihood (LL) indiquent que l'attraction entre les deux pronoms dans le premier corpus est statistiquement très significative.

Bien que le décalage entre les deux œuvres soit impressionnant, il ne nous renseigne pas sur le statut qualitatif que les items cherchés sont susceptibles d'y avoir. Voilà pourquoi nous calculons le « profil combinatoire » (BLUMENTHAL 2012) de *je* et de *nous*, c'est-à-dire le classement des cooccurrents spécifiques que l'on trouve dans les contextes gauche et droit de ces deux pronoms. Pour ce faire, il faut établir des lexicogrammes qui permettent de relever tous les items (*graphiques 1 à 4*). A partir de ces lexicogrammes, nous allons relever les verbes afin d'approfondir leur rôle lors de l'analyse qualitative (réunis pour la commodité de la lecture dans *tableau 2*).

3. Analyse qualitative

Cette section proposera des hypothèses par rapport au fonctionnement discursif des cooccurrents spécifiques et leur évaluation. En effet, à travers le classement de certaines cooccurrences semblent transparaître des cas de figure particuliers, par exemple en ce qui concerne les guillemets comme cooccurrent spécifique à gauche dans *Wikipédia*. Dans le contexte gauche de *je* et de *nous*, ces derniers sont susceptibles d'introduire des citations dans le discours. En revanche, les guillemets ne figurent pas parmi les cooccurrences à gauche recensées pour *l'Encyclopédie*. Une première hypothèse serait donc que dans la *Wikipédia*, les deux pronoms pivots apparaissent très fréquemment dans des citations et ne concernent donc pas immédiatement le locuteur de l'article en question. Une deuxième hypothèse est que le locuteur *je* ou *nous* de l'*Encyclopédie* intervient souvent dans les articles en se servant des *verba dicendi*. Voici l'extraction de leurs cooccurrents spécifiques verbaux :

2a)	*Encyclopédie*		2b)	*Wikipédia*	
dire		avoir	dire		avoir
avouer		croire|croître	déclarer		aimer
aimer	**je**	dire	répondre	**je**	vouloir
voici		vouloir	penser		penser
répondre		aller|voir	aimer		croire|croître

2c)	*Encyclopédie*		2d)	*Wikipédia*	
exciter		avoir	déclarer		avoir
mériter		apprendre	dire		connaître
sembler	**nous**	venir	vouloir	**nous**	pouvoir
suffire		parler	pouvoir		savoir
pouvoir		dire	prier		apprendre

TABLE 2 – Cooccurents spécifiques verbaux

Certaines combinaisons comme *je+croire|croître*[5] ou *je+dire* (2a) ou *nous+parler*, *nous+dire*, ou encore la possibilité *nous+venir+dire* (2c) y sont très fréquentes. En outre, ces *verba dicendi* sont probablement souvent dans des constructions enchâssantes du type *je dis que j'ai...*, ce qui expliquerait le fait que le cooccurrent gauche le plus spécifique de *je* dans l'*Encyclopédie* est en effet *je* (cf. graphique 1). Par ailleurs, la fréquence élevée de *nous+apprendre* (2c), peut soutenir l'hypothèse que le locuteur *nous* de l'*Encyclopédie* tente souvent d'inclure le lectorat dans son discours.

Ces hypothèses - qui ne constituent que des cas exemplaires parmi d'autres et dont nous n'avons donné qu'un bref aperçu pour le moment - seront soumises à un triage à la main, c'est-à-dire à une analyse qualitative de toutes leurs occurrences au sein du concordancier. Ainsi pourra-t-on non seulement dégager les sens et les fonctionnements discursifs des deux pronoms dans les deux corpus, mais aussi procéder à une comparaison. Par exemple, la combinaison de *nous+apprendre* est à la fois répandue dans l'*Encyclopédie* (2c) et dans la *Wikipédia* (2d). Mais ce n'est que l'analyse qualitative des occurrences qui pourra montrer si *nous+apprendre* « fonctionne » de la même manière dans les deux corpus, en particulier en ce qui concerne l'inclusion du lectorat dans le discours.

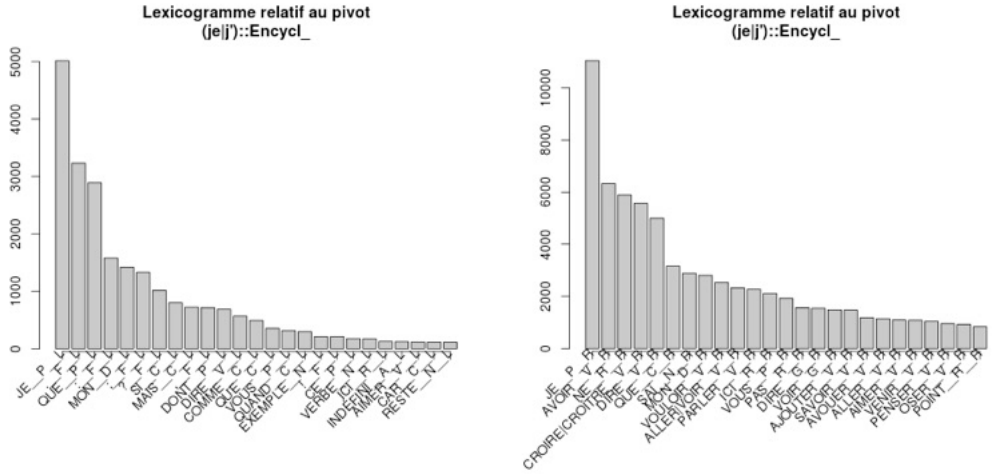
4 Mise en contexte des résultats

Ce panorama succinct aura montré que le locuteur *je* et *nous* est plus courant dans l'œuvre du XVIII[e] siècle que dans celle de nos jours. Dans la foulée, plusieurs questions de posent : comment se change le rapport entre le locuteur et les contenus qu'il met en discours si la première personne n'apparaît plus guère dans la *Wikipédia* (excepté dans des citations) ? Quels sont les impacts sur l'organisation textuelle et son évolution ? Comment interpréter les résultats de l'analyse dans le contexte du changement des rapports entre langue et transmission du savoir dans le temps ? Notre analyse croisée des analyses quantitative et qualitative permet un regard novateur sur la problématique du rapport entre le locuteur et son discours ; la dernière partie de notre communication cherchera à esquisser des possibles réponses à ces questions et posera le point de départ pour une discussion.

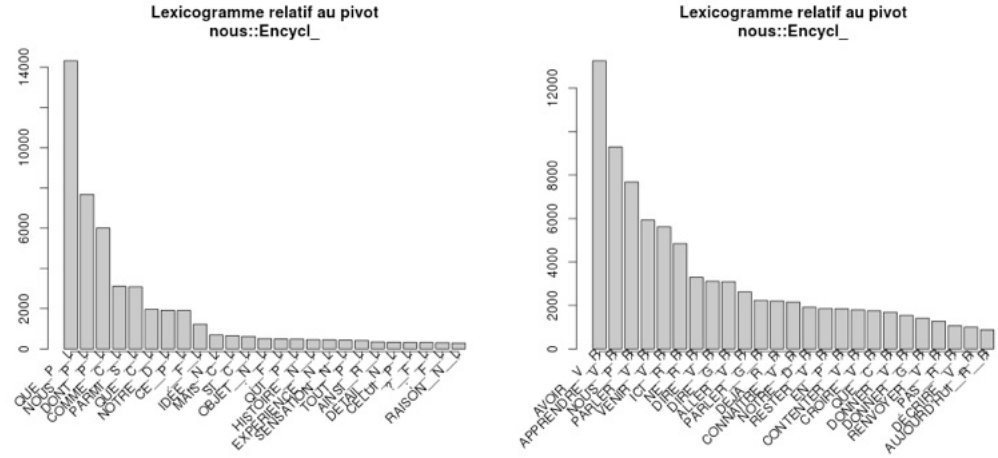
5. Comme la première personne du verbe *croître* semble être rare, on peut partir du principe que la majorité de la forme *je crois* est à attribuer au verbe *croire*.

5. Graphiques

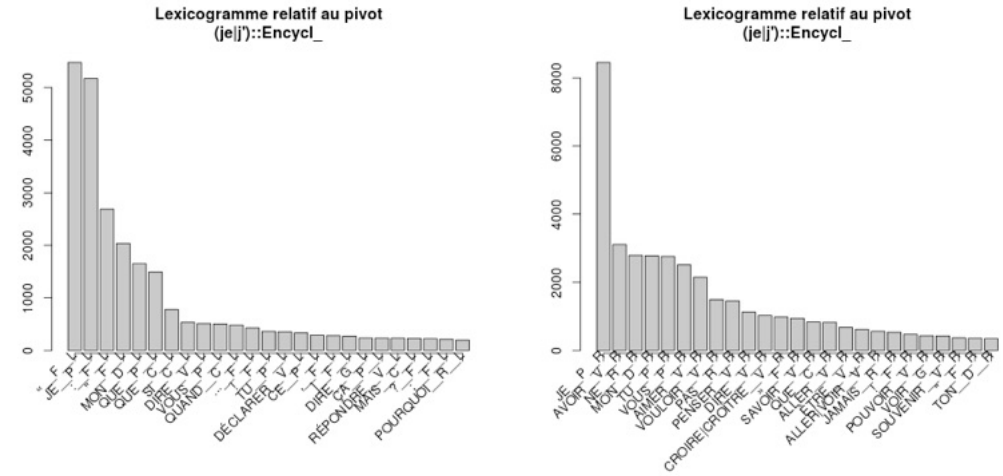
Graphique 1 : lexicogrammes de je *dans l'Encyclopédie (cooccurrents spécifiques à gauche et à droite)*



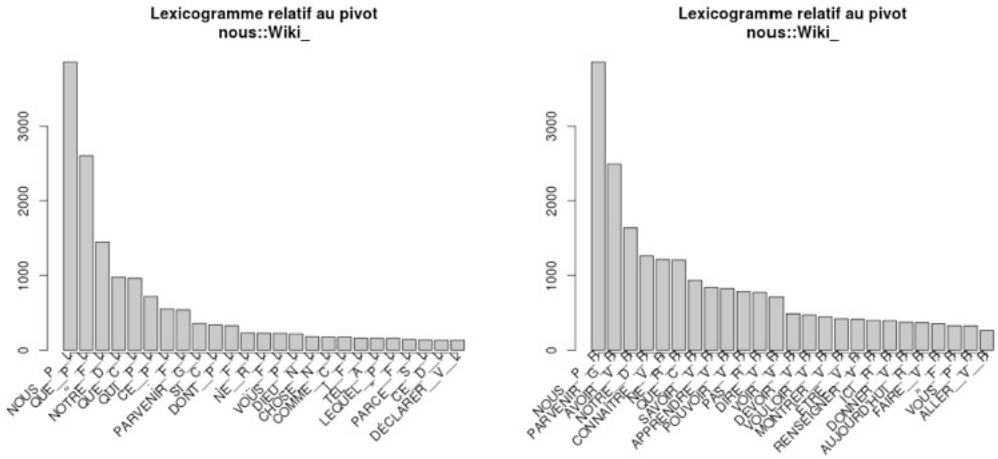
Graphique 2 : lexicogrammes de nous *dans l'Encyclopédie (cooccurrents spécifiques à gauche et à droite)*



Graphique 3 : lexicogrammes de je *dans Wikipédia (cooccurrents spécifiques à gauche et à droite)*



Graphique 4 : lexicogrammes de nous *dans Wikipédia (cooccurrents spécifiques à gauche et à droite)*



Références bibliographiques

Blumenthal, P. (2012). Méthodes statistiques en lexicologie contrastive. In : Begioni, L. et Bracquenier, C. (s.l.d.). *Sémantique et lexicologie des langues d'Europe*. Rennes : Presses Universitaires de Rennes, p. 113-128.

Diwersy, S. (2014). La plateforme Primestat. BTLC et l'exploitation lexico-statistique de corpus diachroniques. *Communication présentée à la Journée d'études organisée à l'Institut des Sciences du Langage et de la Communication*, Université de Neuchâtel.

Evert, S. (2008). Corpora and collocations. In : Lüdeling, A. et Kytö, M. (éds). *Corpus Linguistics. An International Handbook*. Berlin : Mouton de Gruyter, p. 1212-1248.

Gries, S. (2013). *Statistics for Linguists with R. A practical introduction*. Berlin, Boston : De Gruyter.

Groult, M. (2011). *Savoir et Matières. Pensée scientifique et théorie de la connaissance de l'*Encyclopédie *à l'*Encyclopédie méthodique. Paris : CNRS Éditions.

Constitution d'un corpus d'arabe tunisien parlé à Orléans

Youssra Ben Ahmed
Laboratoire Ligérien de Linguistique (UMR 7270)
ben.ahmed.yossra@gmail.com

Etape-clé d'un travail de recherche, la constitution de corpus implique nombre de choix théoriques et techniques, dont l'explicitation permet de mesurer la pertinence des analyses linguistiques et leur probité. Nous nous proposons dans cette communication, qui s'inscrit dans le cadre d'une recherche doctorale en cours, ayant pour objet l'expression du futur en français et en arabe tunisien parlés, d'exposer les principaux choix opérés lors de la constitution du corpus de l'arabe tunisien (désormais AT) et de son annotation.

1 Comparabilité avec ESLO

Après avoir pendant longtemps privilégié les exemples fabriqués ou attestés mais essentiellement écrits et littéraires (pour des raisons à la fois épistémologiques et pratiques), le domaine de la temporalité bénéficie depuis une décennie (si on fait abstraction de quelques précurseurs) de l'avènement de corpus oraux. La prise en compte des données orales y est vue comme une occasion de renouveler profondément les problématiques liées à la temporalité et revigorer les descriptions, qui pouvaient ainsi échapper à l'intuition et à la norme (cf. Abouda 2015).

Le choix de travailler sur des données orales justifié, s'est posée la question du corpus précis sur lequel porteraient nos investigations. Notre choix s'est porté, pour la partie française, sur des données puisées dans les Enquêtes Socio-Linguistiques à Orléans (désormais ESLO). En plus d'une taille conséquente (à ce jour, environ 7 millions de mots) et d'une diversité de genres (conférences universitaires, entretiens, repas en famille ou entre amis...), ESLO offre un avantage essentiel dans le domaine de la temporalité où le pertinence des analyses est tributaire d'une bonne prise en compte de la situation de communication : il contient des données situées, enrichies par des métadonnées qui renseignent sur la situation de communication et précisent pour chaque locuteur son profil en termes d'âge, de sexe et de catégorie socio-professionnelle.

S'agissant d'une recherche contrastive, les choix présidant à la constitution du corpus de l'AT ont été en grande partie dictés par la recherche de la plus grande comparabilité possible avec le corpus ESLO. A commencer par le lieu de l'enquête : même si Orléans constitue sans doute un « non-choix » (Abouda & Baude, 2009 : 133), il nous a paru prudent de respecter cette unité de lieu, d'autant que notre corpus de l'AT a pu ainsi intégrer, pour être partagé, la base de données constituée par le programme « Langues en Contact à Orléans » (LCO) en bénéficiant des moyens et de l'expérience accumulée dans ce cadre. Il va de soi que ce choix, qui a écarté l'autre possibilité un temps envisagée de mener une enquête comparable dans une ville tunisienne, présentait quelques inconvénients : il était particulièrement long et difficile de constituer un corpus suffisamment grand et équilibré.

Malgré ces difficultés, le corpus que nous avons constitué contient 17h d'enregistrement, auprès des tunisiens natifs résidant à Orléans et qui font partie de l'association des Tunisiens du Loiret.

Notre corpus nous semble suffisant pour les investigations envisagées, et parvient à capter une certaine diversité de locuteurs (en fonction des variables d'âge, de sexe, de cadre socioprofessionnelle (CSP), de lieu de naissance et de temps de présence en France), ce qui permet d' « améliorer [sa] représentativité» [1]. Respectant les procédures suivies par ESLO, nécessaires pour rendre le corpus disponible, nous avons procédé à une documentation précise de nos données [2].

En ce qui concerne le mode de recueil des données, nous avons privilégié l'entretien en face-à-face, « situation certes très formelle, mais qui avait l'avantage d'être (...) contrôlable » (Abouda & Baude, 2009 : 134).

Aussi, dans le même souci de comparabilité, nous avons réalisé un questionnaire basé sur les six thèmes retenus par ESLO (logement/Orléans, travail, loisirs, questions évaluatives sur Orléans, langue,

1. Cf. Habert 2000.

2. Pour chaque locuteur, nous avons réalisé une fiche d'information récapitulant l'âge, le sexe, le niveau scolaire... complétée par des indications sur l'enregistrement (n°, type (situation de parole), participant(s), lieu, date et durée de l'enregistrement, situation d'enregistrement...)

recette), afin de faire parler les locuteurs, en ciblant les contextes propices à l'émergence des formes verbales au futur.

2 Annotations

2.1. Transcription

Après la collecte des données, s'est posée la question de leur transcription. Cette étape a soulevé plusieurs interrogations depuis le système graphique jusqu'aux outils de transcription, en passant par le mode de transcription et les conventions adoptées.

Les travaux sur l'arabe tunisien, peu nombreux, hésitent entre les deux systèmes graphiques, i.e. latin et arabe. Le choix de l'un ou l'autre système est dicté par de nombreux paramètres, allant de la tradition du champ, jusqu'aux préférences idéologiques, en passant par la facilité technique. C'est précisément pour cette dernière raison, que nous avons opté pour une transcription avec une graphie latine, qui aura également l'avantage de fournir un corpus partageable et facilement lisible par les non-natifs, i.e. toutes personnes ne maîtrisant pas l'arabe tunisien.

Quant au mode de transcription, entre plusieurs types de notation (phonétique, phonologique, morphologique et usuelle), nous avons opté pour une notation orthographique, même si ce choix n'est pas, ainsi que nous tenterons de le montrer, sans poser de nombreux problèmes en l'absence d'un standard stabilisé. Nous exposerons les conventions choisies en les comparant aux principales conventions proposées.

En ce qui concerne l'outil de transcription, nous avons choisi TRANSCRIBER [3] , un logiciel d'aide à la transcription manuelle de fichiers audio qui permet de transcrire de nombreuses langues y compris non européennes. Nous tenterons de montrer les avantages qu'offre cet outil.

2.2. Annotation

En l'absence d'un étiqueteur morpho-syntaxique, pour exploiter le corpus constitué, nous avons été amené à baliser manuellement les occurrences de futur dans un fichier Transcriber ([lex=FUT-]...[-lex=FUT]). Les 3657 occurrences de futur ainsi identifiées ont été extraites grâce au logiciel d'analyse textométrique TXM [4], et exportées dans un tableau CSV, afin d'y être annotées. Chacune des occurrences du futur a ainsi été sous-spécifiée pour un certain nombre de traits morphosyntaxiques (formes du futur, aspect grammatical, personne, etc.) et sémantiques (types d'emploi : temporel ou modal), dont nous présenterons brièvement la structuration arborescente.

La réinjection sous TXM après annotation des occurrences identifiées permettra une analyse linguistique fine croisant approches qualitative et quantitative, à partir de la création des sous corpus et des partitions, afin d'établir des liens entre les statistiques obtenus et les contenus de nos données.

3 Références bibliographiques

Abouda, L. (2015). *Syntaxe et Sémantique en corpus. Du temps et de la modalité en français oral*, mémoire HDR, Université d'Orléans.

Abouda, L. & Baude, O. (2006). Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des ESLO, *in* F. Rastier, M. Ballabriga (dir.), *Corpus en Lettres et Sciences sociales — Des documents numériques à l'interprétation*, actes du XXVII colloque d'Albi, Langages et signification, publiés par C. Duteil-Mougel et B. Foulquié.

Abouda, L. & Baude, O. (2009). Du français fondamental aux Eslo, *Les Cahiers de Linguistique de Louvain*, 33, 2.

Baude, O. (coord.) (2006). *Corpus oraux, Guide des bonnes pratiques.* CNRS éditions et P.U.O.

Baude, O. (2008). Le droit de la parole, dans Bilger, Mireille (éd.). *Données orales : les enjeux de la transcription.* Perpignan. PUP.pp. 24-33

Benjelloun, S. (2002). Une double graphie, latine et arabe, pour enseigner l'arabe marocain, in : D. Caubet, S. Chaker, J. Sibille (éds), *Codification des langues de France*, pp. 331-340, L'Harmattan, Paris.

3. Téléchargeable sur : http://www.ldc.upenn.edu/mirror.Transcriber/
4. http://textometrie.ens-lyon.fr/

Bergounioux, G. (dir.) (1992). Enquêtes, Corpus et Témoins, *Langue Française* 93.

Bilger, M. (2008). Les enjeux des choix orthographiques dans Bilger, Mireille (éd.) *Données orales – Les enjeux de la transcription*. Perpignan. PUP. pp. 248-257.

Blanche-Benveniste, C. & Jeanjean, C. (1987). *Le français parlé : transcription et édition*, Paris, Didier-Erudition.

Bourdieu P. (2003). (sous la direction de) *La misère du monde*, Paris, Seuil – Collection Point

Caubet, D. (1999). Arabe maghrébin : passage à l'écrit et institutions, In *Faits de Langues*, vol. 7, n° 13, pp. 235-244.

Caubet, D. (2002). Arabe maghrébin, langue de France : entre deux graphies, in : D. Caubet, S. Chaker, J. Sibille (éds), *Codification des langues de France*, p.331-340, L'Harmattan, Paris, 2002 Cerquiglini, B. (1999), *Les langues de la France*, rapport aux ministres de l'Éducation nationale et de la Culture et de la Communication. (en ligne : http://www.dglf.culture.gouv.fr/lang-reg/rapport_cerquiglini/langues-france.html)

Gadet, F. (2000). Derrière les problèmes méthodologiques du recueil des données, dans M. Bilger (dir.), Linguistique sur corpus, Presses Universitaires de Perpignan.

Gadet, F. (2008), L'oreille et l'oeil à l'écoute du social, dans Bilger, Mireille (éd.). Données orales : les enjeux de la transcription. Perpignan. PUP. 35-47.

Habert, B., Nazarenko, A. & Salem, A. (1997). Les linguistiques de corpus, Paris, A. Colin.

Habert B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment ?, dans M. Bilger (dir.), Linguistique sur corpus, Presses universitaires de Perpignan.

Maurer, B. (1999). Quelles méthodes d'enquête sont effectivement employées aujourd'hui en sociolinguistique, dans L.-J. Calvet et P. Dumont (dir.), L'enquête sociolinguistique, L'Harmattan.

Mondada, L. (2008). La transcription dans la perspective de la linguistique interactionnelle, dans Bilger, Mireille (éd.). Données orales : les enjeux de la transcription. Perpignan. PUP. pp. 78-109.

Démonstrations

CLAPI : des corpus écologiques d'interactions et des outils de requêtes

Carole Etienne , Émilie Jouin-Chardon : Laboratoire ICAR - équipe LIS
Adresse URL : http://clapi.ish-lyon.cnrs.fr

CLAPI , Corpus de LAngue Parlée en Interaction, est une plateforme comprenant une banque de données multimédia de 60 corpus enregistrés en situation réelle et dans des contextes variés : interactions privées, professionnelles, institutionnelles, commerciales, didactiques, médicales ... et un ensemble d'outils de requêtes basés sur le lexique (concordancier, co-occurrences, répétitions, ...) et/ou sur des phénomènes propres à l'oral (chevauchement, pause, ...). Actuellement, les requêtes portent sur 63h de données majoritairement vidéo dont 46h sont téléchargeables.

Varitext : une plate-forme pour l'analyse outillée des variétés nationales du français et de l'espagnol

Sascha Diwersy : UMR-CNRS PRAXILING Montpellier 3
Adresse URL : http://syrah.uni-koeln.de/varitext/

Le projet Varitext s'adresse à tous les enseignants-chercheurs qui s'intéressent à l'analyse outillée des variétés nationales de différentes langues telles que le français et l'espagnol. Il met à la disposition de la communauté scientifique une plate-forme d'analyse qui permet l'extraction de concordances, le calcul de spécificités fréquentielles et de cooccurrences lexico-syntaxiques. A présent, les corpus textuels réunis dans la base Varitext couvrent une aire géographique importante des espaces francophone et hispanophone à l'échelle mondiale.

Le site REDAC, Ressources développées à CLLE-ERSS

Cécile Fabre : CLLE, équipe ERSS
Adresse URL : http://redac.univ-tlse2.fr/

REDAC est un site web qui regroupe les ressources linguistiques développées au laboratoire CLLE-ERSS : corpus, ressources lexicales, applications. Ces ressources sont documentées et peuvent être consultées ou téléchargées à partir du site. La présentation se focalisera sur certaines des ressources disponibles, et en particulier sur des corpus consultables en ligne : ParCoLab, corpus parallèle anglais-français-serbe et BaTelOc, base textuelle en langue occitane.

ESLO : un grand corpus oral accessible

Linda Hriba : Laboratoire Ligérien de Linguistique - Equipe ESLO
Adresse URL : http://eslo.huma-num.fr/

Il s'agira, dans cette démonstration, de présenter les principales caractéristiques du corpus ESLO (Enquêtes Sociolinguistiques à Orléans) et la façon dont nous avons conçu son accessibilité, à la fois au travers d'une application spécifique mais également dans l'optique d'un signalement et d'une diffusion plus large sur différentes plateformes (Cocoon, Ortolang). L'interopérabilité des (méta)données

devient alors l'enjeu premier d'un corpus, dont la vocation est d'être partagé et réutilisé, mais ouvre aussi des perspectives sur les apports du linked open data.

PHuN2.0 : Une plateforme de transcription et d'expérimentation

Anne Vikhrova : LIDILEM

Adresse URL : arc.espace-transcription.org ou bien www.espace-transcription.org (2 versions : une d'expérimentation et l'autre de production)

La plateforme PHuN2.0 a été conçue et développée dans le cadre d'une thèse qui a pour objectif d'évaluer l'efficacité de la méthode de crowdsourcing pour la transcription de manuscrits. À partir des retours d'utilisation et des productions de transcripteurs non-experts nous cherchons à évaluer l'apport potentiel du public aux communautés de chercheurs.