Actes des 10èmes Journées Internationales de la Linguistique de corpus

26-28 novembre 2019

Grenoble, France

Comités

Comité scientifique

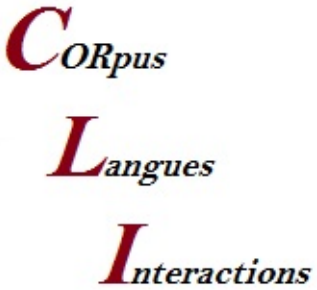
ALEKSANDROVA Tatiana (Université Grenoble Alpes, LIDILEM)
ANTOINE Jean-Yves (Université de Tours, LI)
ANTONIADIS Georges (Université Grenoble Alpes, LIDILEM)
BOULTON Alex (Université de Lorraine, ATILF)
BOUTET Dominique (Université de Rouen, DySoLa)
CARTER-THOMAS Shirley (Institut Mines- Télécom, LaTTiCe)
CAVALLA Cristelle (Université Sorbonne Nouvelle, DILTEC)
DE GIOVANNI Cosimo (Université de Cagliari, Italie)
DIWERSY Sascha (Université Paul Valéry, Praxiling)
DOSTIE Gaétane (Université de Sherbrooke, CANADA)
DUGUA Céline (Université d'Orléans, LLL)
ESKHOL-TARAVELLA Iris (Université d'Orléans, LLL)
ESPERANCA-RODIER Emmanuelle (Université Grenoble Alpes, LIG)
ESTEVE Isabelle (Université Grenoble Alpes, LIDILEM)
ETIENNE Carole (ICAR, CNRS)
FABRE Cécile (Université Toulouse Jean-Jaurès, CLLE-ERSS)
FREROT Cécile (Université Grenoble Alpes, ILCEA4)
FONTENELLE Thierry (Translation Centre for the Corpora of the European Union)
GARDELLE Laure (Université Grenoble Alpes, LIDILEM)
GASIGLIA Nathalie (Université de Lille3, STL)
GRANGER Sylviane (Université Catholique de Louvain, Belgique)
GRESLOU Elisabeth (Université Grenoble Alpes, Litt&Arts)
GROSSMANN Françis (Université Grenoble Alpes, LIDILEM)
HARTWELL Laura (Université Toulouse Capitole, LAIRDIL)
HO-DAC L.-Mai (Université Toulouse Jean-Jaurès, CLLE-ERSS)
HUNSTON Susan (University of Birmingham)
JACKIEWICZ Agata (Université Paul Valéry, Praxiling)
JACQUES Marie-Paule (Université Grenoble-Alpes, Lidilem)
KRAIF Olivier (Université Grenoble-Alpes, Lidilem)
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)
OUNOUGHI Samia (Université Grenoble Alpes, LIDILEM)
PARISSE Christophe (MoDyCo, Inserm)
PECMAN Mojca (Université Paris Diderot - Paris 7, CLILLAC-ARP)
PIERREL Jean-Marie (Université de Lorraine, ATILF)
PONTON Claude (Université Grenoble Alpes, LIDILEM)

POUDAT Céline (Université de Nice, BCL)
REBEYROLLE Josette (Université Toulouse Jean-Jaurès, CLLE-ERSS)
ROSSATO Solange (Université Grenoble Alpes, LIG)
ROUSSET Isabelle (Université Grenoble Alpes, LIDILEM)
SIMON Anne-Catherine (Université catholique de Louvain, Belgique)
SORBA Julie (Université Grenoble Alpes, LIDILEM)
TUTIN Agnès (Université Grenoble-Alpes, LIDILEM)
VIGIER Denis (Université Lumière Lyon II, ICAR)
WILLIAMS Geoffrey (Université de Bretagne-Sud, Litt& Arts)
ZAMPA Virginie (Université Grenoble Alpes, LIDILEM)

Comité d'organisation

Présidente
Marie-Paule Jacques (Université Grenoble Alpes, LIDILEM)
Membres
Tatiana Aleksandrova (Université de Grenoble-Alpes-Alpes, LIDILEM)
Emmanuelle Esperança-Rodier (Université de Grenoble-Alpes-Alpes, LIG)
Cécile Frérot (Université de Grenoble-Alpes, ILCEA4)
Laure Gardelle (Université de Grenoble-Alpes, LIDILEM)
Elisabeth Grelou (Université de Grenoble-Alpes, Litt&Arts)
Marie-Paule Jacques (Université Grenoble Alpes, LIDILEM)
Olivier Kraif (Université de Grenoble-Alpes, LIDILEM)
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)
Geoffrey Williams (Université de Grenoble-Alpes, Litt&Arts)
Virginie Zampa (Université de Grenoble-Alpes, LIDILEM)

Partenaires



Sommaire

Plénière

What we (think) we know about academic writing, and what the corpus evidence is

Hilary Nesi
Faculty Research Centre for Arts, Memory and Communities, Coventry University

Academic discourse has been investigated from many different perspectives, in terms of its lexis, syntax, and rhetorical structure, and with reference to levels of writer/reader expertise, and national and disciplinary cultures. Our understanding of the nature of written academic texts has improved enormously in recent years thanks to corpus linguistic investigations, and older claims are continually being revised as we discover more about the writing produced in specific contexts, for specific readerships and purposes. In this talk I will examine the long-held belief that English academic writing is characteristically 'elaborate', with plenty of subordinate clauses, and compare this with the more recent claim that it is characteristically 'dense', with less subordination and longer nounphrases (see, for example, Biber, Gray & Poonpon, 2011; Parkinson & Musgrave, 2014). Our findings from multidimensional analyses of the BAWE corpus show that elaborate, clausally complex academic writing is more likely to occur in some contexts, and dense, phrasally complex academic writing is more likely to occur in others. Moreover there seem to be two very different types of dense academic writing (one associated with the sciences and the other with the social sciences) and two very different types of elaborate academic writing (one associated with the humanities and the other with academic writing for non-expert readers). I will illustrate the talk with corpus evidence and examples of these different writing styles, and consider the implications of these findings for the teaching of English academic writing.

Présentations orales

Session 1.A.
Corpus et développement langagier

Constituer un corpus d'interactions spontanées et induites : quels apports pour l'étude du développement langagier des enfants prématurés ?

Tiphanie Bertin [1], Caroline Masson [1] et Marine Le Mené [1,2].
[1]EA CLESTHIA, Université Sorbonne Nouvelle Paris 3
[2]LiLPa - Université de Strasbourg
{tiphanie.bertin@sorbonne-nouvelle.fr, caroline.masson@sorbonne-nouvelle.fr, lemeneguigoures@unistra.fr

1 Introduction

Les études menées sur le développement langagier des enfants tout-venant présentent des méthodologies variées, allant de l'expérimentation au recueil de productions langagières induites ou spontanées, longitudinales ou transversales (Canut, Vertalier, 2008 ; Karmiloff, Karmiloff-Smith, 2003). Pour une description plus exhaustive des capacités communicatives et langagières de l'enfant, ces méthodes peuvent aussi parfois être utilisées de façon combinée. Cette complémentarité entre données induites et données spontanées paraît d'autant plus importante lorsqu'il s'agit de rendre compte des compétences d'enfants pouvant présenter des atypies de développement (troubles du spectre autistique, troubles spécifiques du développement du langage oral…(da Silva Genest, Masson, 2019), ou encore à risque de présenter des spécificités de développement, comme les enfants prématurés.

Ainsi, les enfants nés prématurément présentent souvent des décalages développementaux à différents niveaux, et notamment au niveau langagier (Crunelle *et al.*, 2003 ; Grooteclaes *et al.*, 2010 ; Kern, Gayraud, 2007 ; Le Normand *et al.*, 2000 ; Nazzi, 2015 pour une synthèse). Ils sont plus à risque de présenter ensuite des retards de langage (Sansavini *et al.*, 2010 ; Nguyen *et al.*, 2018). Des liens possibles sont envisagés avec des facteurs neurologiques ou cognitifs dans le but de réfléchir à un meilleur accompagnement en charge de ces enfants dès leur naissance. L'hypothèse d'un manque d'exposition au langage durant l'hospitalisation est également avancée dans certaines études (Rand, Lahav, 2014 ; Vohr, 2014). En effet, une naissance prématurée n'est pas sans conséquence sur la relation et les interactions entre le bébé et ses parents (Charavel, 2000 ; Muller Nix *et al.*, 2009). Outre les spécificités de situation (bébé en couveuse, sous assistance respiratoire et/ou alimentaire), le comportement interactionnel du bébé prématuré peut également présenter des spécificités (agitation, fuite du regard, hypertonie) (Lejeune, Gentaz, 2015) qui peuvent induire des difficultés pour les parents à entrer en interaction avec lui. Malgré ces difficultés, certaines études ont mis en évidence des compétences communicatives précoces (expression du visage et regards) chez les bébés prématurés (Tremblay-Levau *et al.*, 1999).

Le lien entre développement communicatif (échange de regards, mimiques, gestes,…) et développement langagier (lexique, morpho-syntaxe, aspects sémantico-pragmatiques…) n'est plus à démontrer chez les enfants tout-venant (Bruner, 1983, 1987 ; Tomasello, 2003). Au vu des compétences communicatives précoces des bébés prématurés et des décalages possibles qu'ils

peuvent présenter au niveau du développement langagier, nous nous interrogeons sur les particularités du lien, de la continuité entre ces deux développements chez ces enfants.

Le projet PREMILA, que nous menons actuellement, vise à décrire l'évolution des compétences communicatives d'enfants prématurés, en lien avec leur développement langagier et avec les interactions dont ils ont bénéficié et dont ils bénéficient au cours de leur développement. La question que nous nous posons est de savoir si le décalage développemental constaté chez ces enfants concerne autant le développement communicatif que le développement langagier. Comment pourrait-on s'appuyer sur les compétences communicatives précoces pour soutenir le développement langagier de ces enfants ? Quel est le rôle des interactions précoces parents-enfants dans ce processus ?

Dans cette perspective, nous nous sommes interrogées sur le type de données à recueillir et à analyser. La dimension longitudinale nous semble fondamentale pour saisir ces processus mais qu'en est-il de la nature des données : spontanées, induites ? Comment rendre compte des compétences communicatives et langagières de ces enfants ? La complémentarité de données d'interaction spontanées et induites nous semble alors primordiale pour mieux saisir les compétences de ces enfants.

Nous proposons dans cette présentation de faire le point sur ces questions méthodologiques en présentant l'intérêt, les conditions et les difficultés d'un tel recueil de données langagières. Nous discuterons également des apports et des limites d'un tel corpus, pour la recherche citée et pour les recherches sur le langage de l'enfant en général, ainsi que pour les applications possibles des résultats d'une telle recherche.

2 Corpus et méthodologie

2.1 Corpus : Constituer un corpus longitudinal auprès d'enfants prématurés

La constitution d'un corpus longitudinal implique un suivi régulier des enfants. Notre objectif était de pouvoir suivre ces enfants tous les 3 mois pendant deux ans, de leur 3 mois à leur 24 mois. Sept enfants nés entre 26 et 32 semaines d'aménorrhées sont ainsi suivis depuis octobre 2018. Étant donné la santé fragile de ces enfants, des visites au domicile familial nous ont semblé la meilleure option à proposer aux familles. Nous présenterons les difficultés et les précautions à prendre pour une telle recherche de terrain, liées à la fois à toute recherche de linguistique impliquée et à la particularité des enfants prématurés. Comme pour tout suivi longitudinal, une relation de confiance et durable doit être établie avec l'enfant et sa famille. En outre, ce type de suivi représente un investissement en temps de la part des parents, déjà très pris par les multiples rendez-vous médicaux de suivi de leurs enfants durant les deux premières années.

2.2 Méthodologie : Complémentarité de données induites et spontanées

Évaluer les compétences communicatives et langagières d'un enfant nécessite de mettre en place des situations spécifiques permettant de provoquer, d'induire certains comportements in-

teractionnels sur une courte durée, qui pourraient ne pas se manifester au cours d'une interaction spontanée. Nous avons donc fait le choix de travailler à partir d'un outil spécifique, l'Echelle de la Communication Sociale Précoce (ECSP, Guidetti, Tourrette, 2009), qui propose un matériel spécifique (ensemble d'objets, de jouets, de livres), lié à une liste de situations à mettre en œuvre avec l'enfant (par exemple : remonter un objet mécanique et le poser hors de portée de l'enfant, solliciter un jeu d'échange de balle auprès de l'enfant···). Ces situations permettent d'évaluer les capacités d'attention conjointe, d'interaction sociale et de régulation du comportement. Ces moments d'échanges avec l'enfant sont menés par le chercheur et filmés (45 minutes à 1h par séance). Cependant, même s'il s'agit de situations avec lesquelles l'enfant est familiarisé, elles restent en partie construites et l'enfant peut ou non s'y engager pleinement.

Il nous a donc semblé important de compléter ce recueil de données induites par un recueil de données spontanées. Ainsi, parallèlement à ces situations, les parents sont filmés en interaction avec leur enfant, dans une situation de change/habillage et de jeu/narration (15 à 30 minutes par séance). Ces vidéos permettent de saisir des compétences communicationnelles et linguistiques qui n'apparaissent pas, ou moins, dans le cadre de situations construites avec une personne non familière à l'enfant. Elles présentent également l'intérêt d'illustrer les comportements interactionnels d'étayage parental et les réactions de l'enfant à ces conduites, illustrant ainsi les compétences interactionnelles et dialogiques de la dyade.

3 Résultats : Apports d'un corpus longitudinal de données d'interactions adulte-enfant prématuré

Nous conclurons notre présentation en réfléchissant aux apports d'un tel corpus à l'étude du langage des enfants prématurés et à l'étude du langage de l'enfant en général. Ainsi, notre corpus pourra contribuer à alimenter les bases de données existantes sur le langage de l'enfant. Les suivis longitudinaux des enfants prématurés, observés notamment dans les cohortes EPIPAGE (Etude épidémiologique sur les petits âges gestationnels, INSERM, Unité 1153, équipe EPOPé), sont souvent effectués à partir de tests de langage ou de questionnaires parentaux, afin de donner des repères développementaux. Peu de données d'interactions spontanées et induites entre adulte et jeunes enfants nés prématurément sont disponibles.

Alors que les études menées sur le développement langagier des enfants prématurés relatent surtout les états des décalages présentés par ces enfants, un tel corpus longitudinal permettra d'enrichir ces résultats avec de nouvelles observations sur la façon dont se compensent ces décalages au cours des deux premières années. Nos analyses ont pour objectif de contribuer à mieux saisir les processus d'évolution langagière de ces enfants, aux différents niveaux linguistiques. En outre, le recours à des situations semi-expérimentales et à des situations d'interactions spontanées devrait permettre de compléter les descriptions sur les compétences communicatives de ces enfants. La mise en relation des analyses sur l'évolution de leurs compétences communicatives et langagières nous amènera à réinterroger l'importance du lien entre ces deux aspects développementaux décrits chez les enfants tout-venant : quelle transition de l'un à l'autre ? Quel appui de l'un sur l'autre ? Comment les capacités communicatives des bébés prématurés servent

leur développement langagier ? Est-ce semblable aux processus observés chez les enfants tout-venant ?

Enfin, nous terminerons notre présentation en nous interrogeant sur les intérêts pratiques et applicatifs d'une telle recherche. En premier lieu, alors que les aspects développementaux langagiers de ces enfants sont souvent décrits en termes de manque, de décalages, il nous semble important de valoriser les compétences communicatives et langagières de ces enfants en décrivant leur processus développemental. En second lieu, une meilleure connaissance de ces processus, observés dans le cadre d'interactions adulte-enfant, permettra de mener une réflexion avec les adultes qui entourent ces enfants (orthophonistes, parents, professionnel.le.s de la petite enfance…) sur la position interactionnelle et sur l'apport linguistique le plus pertinent à leur fournir. Cela permettra de compléter les résultats des travaux menés sur la guidance parentale auprès des enfants prématurés (notamment ceux du projet EPILANG, Charkaluk, 2014).

Références bibliographiques

Bruner, J.S. (1983). *Le développement de l'enfant : savoir faire, savoir dire.* Paris : PUF.

Bruner, J .S. (1987). *Comment les enfants apprennent à parler.* Paris : Retz.

Charavel, M. (2000). Évolution de l'attitude des mères d'enfant prématuré et des mères d'enfant à terme en interaction avec leur bébé : une étude éthologique de la naissance à 6 mois. *La Psychiatrie de l'Enfant*, 43.1, 175-206.

Crunelle, D., Le Normand M.T. & Delfosse, M.J. (2003). Langage oral et écrit chez des enfants prématurés : résultats à 7ans et demi. *Folia Phonatrica et Logopaedica*, 55, 115-127.

Da Silva Genest, C., Masson, C. (2019). Corpus et pathologies du langage : du recueil à l'analyse de données pour une linguistique clinique et appliquée. *Corpus* [En ligne], 19.

Grooteclaes, V., Docquier L. & Maillart C. (2010). Le langage spontané des enfants prématurissimes : analyse du langage descriptif et informatif. *Glossa*, 108, 1-17.

Guidetti M. & Tourrette, C. (2009). *Echelle d'évaluation de la Communication Sociale Précoce*, Paris : Eurotests Editions.

Kern, S. & Gayraud, G. (2007). Influence of preterm birth on early lexical and grammatical acquisition. *First Language*, 27.2, 159-173.

Lejeune, F. & Gentaz, E. (2015). *L'enfant prématuré. Développement neurocognitif et affectif.* Paris : Odile Jacob.

Le Normand, M.T., Parisse, C. & Crunelle, D. (2000). Acquisition du langage chez l'enfant à risque biologique et social : le cas des enfants prématurés. *Rééducation orthophonique*, 38.202, 11-38.

Muller Nix C., *et al.* (2009). Prématurité, vécu parental et relations parents/enfant : éléments cliniques et données de recherche. *La psychiatrie de l'enfant*, 52.2, 423-450.

Nazzi, T. (2015). Acquisition du langage chez l'enfant prématuré durant la première année de vie. *Archives de pédiatrie*, 22, 1072-1077.

Nguyen, TN., Spencer-Smith, M., Zannino, D., *et al.* (2018). Developmental trajectory of language from 2 to 13 years in children born very preterm. *Pediatrics*, 141.5.

Rand, K. & Lahav, A. (2014). Impact of the NCIU environment on language deprivation in preterm infants. *Acta Paediatrica*, 103, 243-248.

Sansavini, A., Guarini A., Justice, L., *et al.* (2010). Does preterm birth increase a child's risk for language impairment ?. *Early Human Development*, 86, 765-772.

Tomasello, M. (2003). *Constructing a language : a usage-based theory of language acquisition*, Cambridge, MA : Harvard University Press.

Tremblay-Leveau, H., Lemaitre-Boquet, L., Megan, A., *et al.* (1999). Les actions de communication chez les bébés prématurés à haut risque. *Enfance*, 52.1, 33-42.

Vohr, B. (2014). Speech and language outcomes of very preterm infants. *Seminar in Fetal & Neonatal Medicine,* 19, 78-83.

Recueil non supervisé et traitement d'un corpus oral dense et massif dans une école maternelle : un exemple avec le projet DyLNet

Aurélie Nardy [1], Isabelle Rousset [1], Hélène Bouchet [1], Loïc Liegois [2], Laurence Buson [1], Céline Dugua [3] et Jean-Pierre Chevrot [1].

[1]LIDILEM, Université Grenoble Alpes

[2]LLF et CLILLAC, Université Paris 7

[3]LLL, Université d'Orléans

aurelie.nardy@univ-grenoble-alpes.fr, isabelle.rousset@ univ-grenoble-alpes.fr,

helene.bouchet@univ-grenoble-alpes.fr, loic.liegeois@univ-paris-diderot.fr,

laurence.buson@univ-grenoble-alpes.fr, celine.dugua@univ-orleans.fr, jean-pierre.chevrot@univ-grenoble-alpes.fr

Cette communication a pour objet le recueil et le traitement d' un corpus oral dense et massif mis en place dans le cadre du projet de recherche DyLNet impliquant le laboratoire Lidilem (Univ. Grenoble Alpes) et l'équipe DANTE de INRIA Rhône-Alpes (ENS de Lyon).

1 Présentation du projet DyLNet

Le but général du projet DyLNet[1] (*Language Dynamics, Linguistic Learning, and Sociability at Preschool : Benefits of Wireless Proximity Sensors in Collecting Big Data*) est d'étudier les relations entre socialisation enfantine et développement du langage oral en maternelle. Il se décline en 2 objectifs principaux :

- Analyser le lien entre interactions sociales et développement du langage oral durant les 3 années de scolarisation en maternelle
- Décrire la coévolution entre dynamique des réseaux sociaux (les changements dans les liens sociaux au sein de l'école) et dynamique du langage dans les réseaux (les influences entre individus et la modification de leurs habiletés langagières).

Sa mise en œuvre repose sur une approche interdisciplinaire novatrice combinant travaux sur l'acquisition du langage, sociolinguistique et science des réseaux (domaine des sciences des systèmes complexes qui observe et modélise tous types de systèmes –sociaux, biologiques ou physiques –composés de nombreuses unités interconnectées) ainsi que sur la collecte et le traitement de données massives (Nardy *et al.*, 2016).

Notre approche consiste à suivre pendant 3 ans près de 200 enfants ainsi que tous les intervenants pédagogiques d'une école maternelle socialement mixte (le recueil de données a débuté en 2017). Tous sont équipés une semaine par mois de capteurs sans fil qui enregistrent, toutes les 5 secondes, les proximités entre individus. De plus, les usages langagiers des enfants et des adultes sont enregistrés grâce à des micros intégrés aux capteurs. À plusieurs reprises durant les trois années, chaque enfant passe également une série de tests psycho- et sociolinguistiques. Enfin, leur

1. Ce projet est financé par l'Agence Nationale de la Recherche (réf. ANR-16-CE28-0013).
Site web : `https://dylnet.univ-grenoble-alpes.fr/`

profil social est établi grâce à un questionnaire rempli par leurs familles. Cette collecte aboutira à la constitution d'une base de données caractérisée par un volume important, une certaine variété (parole, contacts sociaux, informations sociodémographiques, tests) et un flux d'entrée rapide résultant de la capacité des capteurs à enregistrer en continu (De Mauro *et al.*, 2016).

Cette démarche nécessite de relever plusieurs défis : prise en compte des aspects éthiques et traitement des données personnelles, développement de matériel *ad hoc*, stockage, traitement et analyse des données. Dans cette communication, nous nous centrerons sur le recueil et le traitement de corpus oraux denses et massifs tels qu'ils ont été appréhendés dans le cadre du projet DyLNet. Ce type de corpus est d'une grande richesse puisqu'il permet, dans le cadre du développement langagier, d'appréhender l'éventail des usages de l'enfant et ceux de son entourage. Ils sont toutefois à ce jour relativement peu nombreux[2] du fait de différentes contraintes matérielles et techniques liées aux modalités d'enregistrement et au traitement des données (Canault *et al.*, 2017).

2 Modalités d'acquisition des enregistrements audio

Un premier défi, dans le cadre de ce projet, a été de mettre en place un système d'acquisition des enregistrements audio qui satisfasse à la fois aux spécificités de notre terrain d'enquête et à nos besoins pour la recherche.

La majorité des participants sont de jeunes enfants (âgés de 2 ans et demi à 6 ans) qui sont enregistrés en continu une semaine par mois lors de leurs différentes activités quotidiennes à l'école (classe, cour de récréation, sport). Le dispositif d'enregistrement devait donc répondre à des exigences d'innocuité, de facilité d'équipement et ne pas gêner ceux qui le portent. Dans le même temps, il devait inclure une capacité de stockage suffisante (24 heures d'enregistrement chaque semaine), satisfaire des critères de qualité des signaux enregistrés et enfin inclure un dispositif d'horodatage.

En partenariat avec une entreprise de la région Auvergne-Rhône-Alpes, nous avons développé des petits boitiers de 58,15 x 50 x 15 mm (hauteur x largeur x profondeur) portés au col grâce à une pince bretelle qui, en plus d'enregistrer toutes les 5 secondes les proximités entre individus, incluent deux microphones sur la face avant.

3 Modalités de traitement des enregistrements audio

Le nombre de participants ainsi que la masse de données recueillies ($\approx$ 30 000 heures/an) implique de relever deux défis liés au traitement des données audio recueillies.

Le premier nécessite de reconnaître la voix du porteur du micro dans un environnement multi-locuteurs et bruité. Dans le cas de jeunes enfants, la tâche est encore plus complexe puisque, contrairement aux adultes, les fréquences fondamentales moyennes des deux sexes ne sont pas

2. Voir toutefois l'impressionnante étude Human Speechome Project (Roy, et al., 2006).

différentes (Busby & Plant, 1995 ; Lee *et al.*, 1999 ; Weinberg & Bennett, 1971). Le second défi implique d'optimiser la tâche de transcription, fastidieuse et chronophage, à la fois en termes temporels et en termes de qualité pour répondre aux objectifs de la recherche.

3.1 Dispositifs de traitement automatique des signaux

Différents dispositifs de pré- et post-traitements automatiques des enregistrements audio ont été mis en œuvre.

Une fois les enregistrements audio extraits des capteurs, un premier traitement est appliqué afin de procéder au découpage de ces derniers en fichiers d'une durée d'une heure maximum. Les fichiers obtenus conservent leur horodatage initial et sont tous alignés sur la même tranche horaire. Pour chacun de ces fichiers, nous disposons de différentes versions : originales et post-traitées.

Concernant les versions stéréo originales, elles se présentent d'une part au format *.wav* (version RAW sur laquelle se font les transcriptions) et d'autre part au format .flac (format de compression sans perte pour le stockage et la conservation).

À partir de la version stéréo originale, un post-traitement automatique des signaux est appliqué par un filtrage spatial (*beamforming*) réalisé sur la base du décalage physique entre les deux micros intégrés à chaque capteur. Il permet la génération de deux versions post-traitées de la version stéréo originale : version MASKED et version CUT. La version MASKED correspond à une version de l'enregistrement dans laquelle seuls les segments audio identifiés par le post-traitement comme émanant du porteur du capteur sont présents, tout en conservant le décours temporel du fichier audio. La version CUT, quant à elle, correspond à l'enchaînement bout à bout des segments audio identifiés par le post-traitement comme émanant du porteur. Ce post-traitement automatique, qui dépend de la position du locuteur par rapport aux micros du capteur, n'est pas parfait (évaluation de sa fiabilité à venir). Il est toutefois précieux pour deux raisons principales.

Premièrement, il est utile lors de la sélection des fichiers audio à transcrire puisque le ratio entre la durée du fichier RAW et celle du fichier CUT correspondant nous permet d'avoir une idée approximative du temps de parole du porteur sur l'heure d'enregistrement.

Deuxièmement, il est utile lors de la phase de transcription pour cibler la voix du porteur. Sur la version MASKED, un script qui segmente le signal en silence/parole est lancé depuis *PRAAT*. À l'issue de ce processus, nous récoltons un fichier *.TextGrid* qui contient les temps de silence et de parole.

Ce fichier est ensuite ouvert dans le logiciel de transcription alignée *ELAN* avec le fichier audio RAW, générant ainsi des bornes automatiques qui indiquent au transcripteur les passages

lors desquels le porteur parle et lui évitent d'avoir à écouter l'intégralité d'un fichier audio d'une heure lorsque le porteur a manifesté seulement quelques prises de parole.

3.2 Procédure de transcription

Conscients de la complexité de l'activité de transcription et de l'attention qu'elle requiert (Baude & Dugua, 2011 ; Hriba *et al.*, 2011), nous avons mis en place une procédure de transcription en 4 phases distinctes :

1. ajustement des bornes temporelles posées automatiquement et codage de l'activité en cours, du(des) interlocuteur(s) et des situations langagières rencontrées ;
2. transcription des paroles ;
3. vérification de certains codages grâce à des requêtes lancées depuis ELAN ;
4. anonymisation des signaux (en conformité avec les engagements éthiques pris auprès de la CNIL et du comité d'éthique qui a validé le protocole).

De plus, afin de rendre la tâche de transcription la moins lourde et la plus fiable possible, nous avons élaborées des procédures –que nous présenterons –pour détecter automatiquement les contextes de négation et de liaisons facultatives pour un codage ultérieur par un expert.

Enfin, à l'instar des pratiques que l'on retrouve dans des corpus tels que la banque de données *VALIBEL* (Bachy *et al.*, 2007) ou le projet *Traitement de Corpus Oraux en Français (TCOF)* (André & Canut, 2010), chaque transcription sera relue par un membre du projet DyLNet.

Références bibliographiques

André, V. & Canut, E. (2010). Mise à disposition de corpus oraux interactifs : le projet TCOF (Traitement de Corpus Oraux en Français) *Pratiques*, 147-148, 35-51.

Bachy, S., Dister, A., Francard, M., Geron, G., Giroul, V., Hambye, P., Simon, A.-C. & Wilmet, R. (2007). Conventions de transcription régissant les corpus de la banque de données VALIBEL. `http://hdl.handle.net/2078.1/165551.`

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

Busby, P. A. & Plant, G. L. (1995). Formant frequency values of vowels produced by preadolescent boys and girls. *Journal of the Acoustical Society of America*, 97, (4), 2603-2607.

Canault, M., Le Normand, M.-T. & Thai Van, H. (2017). LENA (Language ENvironment Analysis System) : un système de reconnaissance automatique de la parole et de l'environnement langagier de l'enfant. *Enfance*, 2, (2), 199-216.

De Mauro, A., Greco, M. & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review*, 65, (3), 122-135.

Hriba, L., Baude, O. & Dugua, C. (2011). Transcrire : la norme, la variation et le linguiste. *Colloque du CerLiCO : Transcrire, écrire, formaliser 2.* Orléans.

Lee, S., Potamianos, A. & Narayanan, S. (1999). Acoustics of children's speech : developmental changes of temporal and spectral parameters. *Journal of the Acoustical Society of America*, 105, 1455-1468.

Nardy, A., Fleury, É., Chevrot, J.-P., Karsai, M., Buson, L., Bianco, M., Rousset, I., Dugua, C., Liégeois, L., Barbu, S., Crespelle, C., Busson, A., Léo, Y., Bouchet, H. & Dai, S. (2016). DyLNet –Language Dynamics, Linguistic Learning, and Sociability at Preschool : Benefits of Wireless Proximity Sensors in Collecting Big Data (`https://dylnet.univ-grenoble-alpes.fr/`). <ANR-16-CE28-0013>, `https://hal.archives-ouvertes.fr/hal-01396652`.

Roy, D., Patel, R., Decamp, P., Kubat , R., Fleischman, M., Roy, B., Mavridis, N., Tellex, S., Salata, A., Guinness, J., Levit, M. & Gorniak, P. (2006). The Human Speechome Project. *Proceedings of the 28th Annual Cognitive Science Conference.*

Weinberg, B. & Bennett, S. (1971). Speaker sex recognition of 5- and 6-year-old children's voices. *Journal of the Acoustical Society of America,* 50, (4), 1210-1213.

L'entrée à l'école maternelle : quel français scolaire, face à quels usages enfantins ?

Laurence Buson et Aurélie Nardy
LIDILEM, Université Grenoble Alpes
laurence.buson@univ-grenoble-alpes.fr, aurelie.nardy@univ-grenoble-alpes.fr

1 Contexte

Comme le reconnait lui-même le gouvernement, « le poids de l'origine sociale sur les performances des élèves est plus fort en France que dans tous les pays de l'OCDE » [1]. Parce que l'école maternelle est la première étape de la scolarisation, il est nécessaire de comprendre comment les enfants de tous les groupes sociaux s'y côtoient, s'y intègrent et s'y adaptent. Le langage oral joue un rôle central dans ce contact précoce avec le monde scolaire : il est à la fois un moyen et un résultat de la socialisation scolaire et la « condition essentielle de la réussite de toutes et de tous » (M.E.N., 2015). Réciproquement, une socialisation scolaire réussie multiplie les opportunités de communication avec les pairs et les adultes en charge des enfants, et renforce les compétences (socio)linguistiques de tous les élèves. Il peut donc s'établir un cercle vertueux –ou au contraire une spirale d'échec –entre sociabilité enfantine, langage oral et apprentissages scolaires.

Dans cet enchaînement, les inégalités sociales jouent un rôle central puisque, dès l'âge de 2 ans, on observe que les enfants issus de familles plus favorisées ont un lexique plus diversifié et formulent des énoncés plus longs que les enfants issus de familles moins favorisées (Le Normand, Parisse & Cohen, 2008 ; Parisse & Le Normand, 2006). En outre, les enfants de tous les milieux ne sont pas tous aussi familiers et n'utilisent pas au même degré les codes linguistiques valorisés par l'école et ses représentant·e·s (Nardy, Chevrot & Barbu, 2013). Ces différences précoces issues de la transmission au sein de la famille (Huttenlocher, Vasilyeva, Waterfall, Vevea & Hedges, 2007 ; Smith, Durham & Richards, 2013) ont suscité de nombreuses recherches qui ont mis en évidence l'influence de la nature (Hoff, 2002, 2003 ; Hoff, Laursen & Tardif, 2002 ; Rowe, 2008) et de la quantité (Hart & Risley, 2003 ; Hoff-Ginsberg, 1994 ; Hoff & Naigles, 2002 ; Rowe, 2008) de discours adressé à l'enfant par ses parents dans les différents milieux sociaux. Ces travaux laissent toutefois dans l'ombre à la fois l'influence des pairs, notamment lorsque la composition du groupe scolaire est socialement mixte (Buson & Billiez, 2009 ; Schechter & Bye, 2007), et l'influence du discours de l'enseignant·e, par exemple en termes d'écart potentiel entre cette parole institutionnelle légitime et les usages des élèves. La scolarisation génère donc une situation de communication inédite, les interactions avec les pairs et les enseignant·e·s (Bowers & Vasilyeva, 2011 ; Huttenlocher, Vasilyeva, Cymerman & Levine, 2002) étant susceptibles de modifier les usages hérités de la famille.

1. Relevé sur le site du gouvernement français, sur une page mise à jour en mai 2017 : `https://www.gouvernement.fr/action/la-lutte-contre-les-inegalites-scolaires`

Le français scolaire constitue une zone peu explorée de la sociolinguistique, avec peu d'études focalisant sur les caractéristiques du discours auquel sont exposés les enfants dès leur entrée à l'école (Bellonie & Guerin, à paraître ; Boutet, 2003 pour les propriétés énonciatives du français scolaire)[2]. C'est pourquoi nous proposons dans cette étude une description (socio)linguistique du français produit par l'enseignante d'une classe de Petite Section de maternelle (désormais PS) afin d'envisager et d'évaluer la distance entre cette variété « modèle » et les usages langagiers des élèves, eux aussi diversifiés, à leur entrée à l'école.

2 Méthodologie

Les données langagières sur lesquelles se fondent les analyses de cette étude ont été recueillies dans le cadre du projet DyLNet[3] (*Language Dynamics, Linguistic Learning, and Sociability at Preschool : Benefits of Wireless Proximity Sensors in Collecting Big Data*) dont l'objectif principal est d'examiner le lien entre interactions sociales et développement du langage oral dans une école maternelle grâce à un suivi longitudinal d'une durée de 3 ans[4]. Ainsi, une semaine par mois, pendant 3 ans, les enfants et adultes d'une école maternelle d'une zone urbaine du département de l'Isère sont équipés de capteurs qui enregistrent à la fois les proximités et les interactions verbales des individus qui les portent. Par ailleurs, des informations sociodémographiques sur les enfants et leur environnement familial ont été recueillies par un questionnaire adressé aux familles.

Dans la présente étude, nous nous intéresserons aux productions verbales de 17 enfants d'une classe de PS et de leur enseignante, enregistrés pendant une semaine, un mois après la rentrée scolaire. À partir d'un corpus d'une durée totale d'une quinzaine d'heures, transcrit et annoté dans ELAN, nous analyserons les prises de paroles des élèves et de leur enseignante, recueillies à la fois sur des temps de classe et en cour de récréation. L'échantillon comprend 5 filles et 13 garçons, qui ont entre 2 ;10 et 3 ;9 (âge moyen de 3 ans 2 mois) et qui sont issus de milieux sociaux contrastés.

3 Analyses

Les analyses porteront principalement sur les usages sociolinguistiques des participant · e · s aux niveaux phonétique/phonologique (suppression optionnelle de /l/ dans le clitique « il(s) », suppression du « u » dans « tu » devant voyelle) et syntaxique (négation, formulation des questions). En outre, nous mettrons en perspective les productions de variables sociolinguistiques relevées en classe avec celles relevées hors de la classe (cour de récréation) et chercherons à savoir si l'interlocuteur (adresses adulte/adulte, adulte/enfant, enfant/adulte, enfant/enfant) module ces productions.

2. Sur la norme et les discours scolaires, généralement dans une perspective didactique, voir notamment Bautier (2007), Bautier & Branca-Rosoff (2002), Bautier & Rayou (2013), Crinon, Marin & Bautier (2008), François (1980), Genouvrier (1972).

3. DyLNet est un projet financé par l'Agence Nationale de la Recherche (réf. ANR-16-CE28-0013). Site web du projet : `https://dylnet.univ-grenoble-alpes.fr/`

4. Le recueil de données est encore en cours.

Enfin, les discours explicites sur la norme, ainsi que les aspects métalinguistiques ou épilinguistiques, reformulations et hétéro-corrections dans le discours de l'enseignante viendront compléter ces données variationnelles afin de mieux comprendre à quelle variété de français et à quels discours sur la langue ces enfants de PS sont exposés en tout début de scolarisation.

4 Prolongements

Ces premières observations recueillies à l'arrivée à l'école maternelle seront par la suite mises en perspective avec les usages sociolinguistiques enfantins en fin d'année scolaire. Par ailleurs, dans le cadre plus global du projet DyLNet, les résultats des analyses de ce corpus de langage oral seront mis en relation avec les réseaux d'interactions sociales des individus afin d'envisager la question de la socialisation horizontale (quelles similitudes et influences réciproques peut-on observer chez des enfants qui interagissent souvent au cours de l'année scolaire ?) et verticale (dans quelle mesure le discours de l'enseignante influence-t-il, voire est-il influencé par les productions enfantines ?) de ces mêmes élèves.

Références bibliographiques

Bautier, É. (2007). Langue et discours : tensions, ambigüités de l'école envers les milieux populaires. *Le français aujourd'hui*, 156, 57-66.

Bautier, É. & Branca-Rosoff, S. (2002). Pratiques linguistiques des élèves en échec scolaire et enseignement. *Ville-École-Intégration Enjeux*, 130, 196-213.

Bautier, É. & Rayou, P. (2013). *Les inégalités d'apprentissage. Programmes, pratiques et malentendus scolaires.* Paris : Presses Universitaires de France.

Bellonie, J.-D. & Guerin, E. (à paraître). La place des pratiques langagières ordinaires des élèves. *Le Français Aujourd'hui.*

Boutet, J. (2003). De l'inégalité dans l'accès au français scolaire. *Le Français Aujourd'hui*, 141, 12-20.

Bowers, E. P. & Vasilyeva, M. (2011). The relation between teacher input and lexical growth of preschoolers. *Applied Psycholinguistics*, 32, (1), 221-241.

Buson, L. & Billiez, J. (2009). Stylistic repertoires and strategies of 10/11 year-old primary school children. *Corela*, 7, (2), `http://corela.edel.univ-poitiers.fr/document.php?id=2246`.

Crinon, J., Marin, B. & Bautier, É. (2008). Quelles situations de travail pour quel apprentissage ? Paroles des élèves, paroles de l'enseignant. In Bucheton, D. (Ed.), *Le développement des gestes professionnels dans l'enseignement du français : un défi pour la recherche et la formation* (pp.123-147). Bruxelles : De Boeck Supérieur.

François, F. (1980). Analyse linguistique, normes scolaires et différenciations socio-culturelles. *Langages*, 59, 25-52.

Genouvrier, É. (1972). Quelle langue parler à l'école ? Propos sur la norme du français. *Langue Française*, 13, 34-51.

Hart, B. & Risley, T. R. (2003). The early catastrophe : the 30 million word gap by age 3. *American Educator*, 27, (1), 4-9.

Hoff-Ginsberg, E. (1994). Influences of mother and child on maternal talkativeness. *Discourse Processes*, 18, 105-117.

Hoff, E. (2002). Causes and consequences of SES-related differences in parent-to-child speech. In Bornstein, M. H. & Bradley, R. H. (Eds.), *Socioeconomic status, parenting and child development* (pp.147-160). Mahwah : Lawrence Erlbaum Associates.

Hoff, E. (2003). The specificity of environmental influence : socioeconomic status affects early vocabulary development via maternal speech. *Child Development,* 74, (5), 1368-1378.

Hoff, E., Laursen, B. & Tardif, T. (2002). Socioeconomic status and parenting. In Bornstein, M. H. (Ed.), *Handbook of parenting* (pp.231-252). Mahwah : Lawrence Erlbaum Associates.

Hoff, E. & Naigles, L. (2002). How children use input to acquire a lexicon. *Child Development,* 73, (2), 418-433.

Huttenlocher, J., Vasilyeva, M., Cymerman, E. & Levine, S. (2002). Language input and child syntax. *Cognitive Psychology*, 45, (3), 337-374.

Huttenlocher, J., Vasilyeva, M., Waterfall, H. R., Vevea, J. L. & Hedges, L. V. (2007). The varieties of speech to young children. *Developmental Psychology,* 43, (5), 1062-1083.

Le Normand, M.-T., Parisse, C. & Cohen, H. (2008). Lexical diversity and productivity in French preschoolers developmental and biosocial aspects by developmental, gender and sociocultural factors. *Clinical Linguistics & Phonetics,* 22, (1), 47-58.

Ministère De L'éducation Nationale, De L'enseignement Supérieur Et De La Recherche, (2015). Annexe - Programme de l'école maternelle. Bulletin officiel de l'Education Nationale, Bulletin officiel spécial °2 du 26 mars 2015.

Nardy, A., Chevrot, J.-P. & Barbu, S. (2013). The acquisition of sociolinguistic variation : looking back and thinking ahead. *Linguistics,* 51, (2), 255-284.

Parisse, C. & Le Normand, M.-T. (2006). Une méthode pour évaluer la production du langage spontané chez l'enfant de 2 à 4 ans. *Glossa*, 97, 20-41.

Rowe, M. L. (2008). Child-directed speech : relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language*, 35, (1), 185-205.

Schechter, C. & Bye, B. (2007). Preliminary evidence for the impact of mixed-income preschools on low-income children's language growth. Early Childhood Research Quarterly, 22, (1), 137-146.

Smith, J., Durham, M. & Richards, H. (2013). The social and linguistic in the acquisition of sociolinguistic norms : caregivers, children, and variation. *Linguistics*, 51, (2), 285-324.

Session 1.B.
Corpus et apprentissage

Serait-il plus facile de faire passer un chameau par le chas d'une aiguille ? Les défis de la conception d'une interface-usager pour l'exploration d'un corpus de français parlé à visée pédagogique en FLE

Christian Surcouf ,
École de français langue étrangère, Faculté des Lettres, Université de Lausanne (Suisse)
christian.surcouf@unil.ch

La compréhension orale d'une langue –à fortiori étrangère –soulève de redoutables défis dans la mesure où l'auditeur doit en temps réel (Cutler 2012 : 304) :

1. segmenter la chaine parlée en phonèmes (entre dix et quinze par seconde selon Levelt 1994 : 90)
2. identifier les mots de l'énoncé (deux à trois par seconde selon Levelt 1989 : 59)
3. agencer ces mots entre eux pour former des phrases morphosyntaxiquement et pragmatiquement cohérentes

Comme le remarque Porcher (1995 : 45) : « la compétence de réception orale est de loin la plus difficile à acquérir et c'est pourtant la plus indispensable. Son absence est anxiogène et place le sujet dans la plus grande "insécurité linguistique" ». Pourtant « l'écoute est [···] quantitativement la plus présente des activités langagières dans le quotidien des individus » (Parpette 2008 : 221), soit environ 50% du temps éveillé d'après les données présentées par Worthington & Fitch-Hauser (2018 : 6, tableau 1-1). Par ailleurs, du point de vue de l'enseignement/apprentissage des langues, « l'histoire des méthodologies [···] montre que, depuis le début du XX^{e} siècle, et de manière continue depuis les années 60, la stratégie consiste à organiser l'apprentissage d'abord largement autour de la maitrise de l'oral, et dans les situations de vie quotidienne » (Parpette 2018 : 22). En somme, la compréhension orale est une des compétences les plus importantes dans l'apprentissage du français langue étrangère, et si Cuq & Gruca (2002 : 154) assurent qu'elle « a retenu toute l'attention dans les années 1970 et a connu un rayonnement particulier avec l'entrée des documents authentiques[1] dans la classe de langue », plus récemment Parpette (2018 : 19) est d'avis que bien que « la notion de document authentique s'est imposée comme un des éléments structurants de l'enseignement du FLE, l'oral échappe à cette intégration dans les outils pédagogiques jusqu'à un niveau avancé ». Un constat analogue apparait chez Vialleton & Lewis (2014 : 312) dans leur analyse de plusieurs manuels de FLE (voir également Debaisieux & Boulton 2007) :

> when it comes to getting students to cope with the complexity of naturally-occurring speech they need to be confronted with that complexity to start being able to make sense of it. [···] such an ecological approach is currently far from the norm in published course materials. (Vialleton & Lewis 2014 : 312)

1. Tout dépend bien sûr du type de « document authentique ». Un roman lu, un journal télévisé peuvent être authentiques sans pour autant préparer l'apprenant à comprendre les pratiques langagières orales du quotidien, qui n'ont que peu à voir avec de tels « documents authentiques ».

Cette situation parait d'autant plus paradoxale que depuis l'avènement de l'approche communicative, les dialogues présentés dans les manuels de FLE mettent en scène des individus évoluant dans un quotidien vraisemblable du point de vue situationnel (par ex. « deux amis discutent de leurs projets de vacances »), mais dont l'activité langagière ne reflète pas cette vraisemblance. Ravazzolo *et al.* (2015 : 111) signalent ainsi « l'écart entre les usages spontanés de la parole en interaction entre locuteurs natifs et ceux présentés dans les dialogues pédagogiques », en effet, « avant d'être oralisé, le dialogue pédagogique est écrit [⋯] s'inscrivant dans une logique de présentation propre à l'écrit », où n'apparaissent, ni hésitations, ni réductions, ni dislocations, etc. Pourtant, dans sa dimension la plus courante, en face-à-face, l'oral spontané se démarque fortement de la lecture ou de la diction soignée des comédiens sollicités pour l'enregistrement des dialogues des manuels de langues :

> the acoustic realization of continuous speech is severely degraded when compared to the maximally distinct isolated utterances often used in laboratory situations. The acoustic cues that accompany words spoken in isolation are often simply not present in fluent speech ; segments and syllables are omitted, vowel color is significantly changed by consonantal environment (Bond & Garnes 1980 : 116)

S'il semble logique de considérer qu'« on ne peut acquérir que ce que les supports permettent d'acquérir », alors l'accent doit être mis sur les documents sonores reflétant de près les pratiques ordinaires des natifs[2] dans la mesure où « leur authenticité accroit la probabilité qu'ils offriront bien à l'apprenant les moyens d'acquérir les savoirs dont il aura besoin pour fonctionner langagièrement en situation "réelle" » (Holec 1990 : 68). Un tel entrainement à l'écoute de documents authentiques évitera les surprises chez l'apprenant[3] : « c'est aussi un choque pour nous des fois d'apprendre que les francophones natifs ne parlent pas comme on a été enseigné dans les manuels et cours de langues » (anglophone de niveau B2 séjournant en milieu francophone). Un autre apprenant, russophone, du même groupe explicite quant à lui l'une des difficultés rencontrées : « j'ai commencé à parler français selon les règles de l'écriture [⋯] J'estime que les enseignants doivent souligner le fait que la structure des phrases diffère à l'écrit et à l'oral ». Il est clair que si les documents sonores des manuels de FLE reflétaient davantage les pratiques quotidiennes des natifs, et si les enseignants y sensibilisaient explicitement leurs apprenants, les difficultés de compréhension seraient moindres. En définitive, en accord avec Vialleton & Lewis :

> Whenever the purpose of a sequence is to teach students how to understand spoken French as it is used by fluent speakers the audio samples provided should accurately reflect the features found in such speech, and the accompanying materials need to provide tools and strategies to help students to achieve the stated purposes. (Vialleton & Lewis 2014 : 312)

Si, en tant que « collection of authentic texts (including transcriptions of spoken data) [⋯] sampled so that they are representative of a particular language or variety of a language, and [⋯] machine-readable », les corpus oraux semblent répondre à de telles demandes pédagogiques, force est de constater qu'« il n'existe pas encore de grand corpus pour le français –et encore

2. À fortiori lorsque les apprenants évoluent dans un milieu francophone.

3. Nous respectons la graphie et la syntaxe de l'apprenant. Ces extraits ont été collectés en fin de semestre dans des commentaires écrits à propos d'exercices de transcription.

moins pour l'oral[4] » (Boulton & Tyne 2014 : 50) et à fortiori à visée originellement pédagogique (mais voir les réutilisations de PFC-EF, Detey *et al.* 2009 ; Clapi-FLE, Ravazzolo *et al.* 2015). Par ailleurs, « les possibilités de recherche sont en fait les mêmes que pour les corpus écrits » (Boulton & Tyne 2014 : 50), faisant du concordancier l'outil central des requêtes.

Forts de ces constats, nous avons créé **F**LORALE (Français Langue ORALE pour le FLE)[5], une base de données du français parlé à finalité pédagogique, permettant aux apprenants d'accéder à une multitude d'exemples sonores d'environ trois secondes illustrant plus de 250 traits langagiers caractéristiques du français ordinaire (c'est-à-dire « la langue de tous les jours » selon Gadet 1996 : V), auquel les apprenants de FLE sont en général peu familiarisés durant leur apprentissage. Si l'élaboration de ce corpus oral informatisé étiqueté à des fins pédagogiques a soulevé de nombreux défis linguistiques, informatiques et didactiques, la conception de l'interface a constitué l'une des plus grandes difficultés. À titre illustratif, nous reproduisons ci-dessous une capture d'écran donnant un aperçu des résultats d'une requête portant sur « Les petits mots ou expressions pour *ne pas* finir la phrase... », et plus particulièrement ici sur « et tout » :

FIG. 1 : Aperçu des résultats de la recherche portant sur « et tout »

Comme il apparait immédiatement, l'affichage ne s'effectue pas sous la forme d'un concordancier, mais de segments incluant le phénomène recherché (ici « et tout »). En effet, tous les segments –écoutables –ont préalablement été découpés, transcrits et alignés sur le signal sonore, et enfin annotés manuellement dans le logiciel Elan[6], qui nous a permis d'inclure les trente strates d'annotations (« tier ») par locuteur dont nous avions besoin, chacune associée à un vocabulaire contrôlé plus ou moins étendu (allant de deux à vingt-cinq étiquettes).

4. Ce qui s'explique aisément : « The earlier development of written corpora is due to the fact that they are easier to compile, since they require only the gathering of written texts in electronic format and the techniques for transforming written text pages to digital format were developed earlier. For speech, on the other hand, we need a much greater effort : it is necessary, at least, to record natural interactions and properly transcribe them » (Raso & Mello 2014 : 2).

5. Librement accessible à l'adresse : `https://florale.unil.ch/`. Les documents proviennent d'entretiens et de documentaires radiophoniques.

6. Max Planck Institute for Psycholinguistics (Nijmegen) disponible ici : `https://tla.mpi.nl/tools/tla-tools/elan/`.

Depuis l'origine (récente) de ce projet, nous avons imaginé la contrainte maximale consistant à permettre la manipulation de l'interface à un apprenant de FLE de niveau A2 dépourvu de connaissances en linguistique et d'expérience d'utilisation d'outils informatiques d'exploration de corpus[7]. En somme, nous avons essayé de créer un « outil de consultation convivial, adapté à un usage d'apprentissage », dont Debaisieux (2009 : 43) déplorait l'absence il y a encore une décennie. La présentation des résultats dans notre interface devait alors répondre aux deux défis suivants :

a. Comment organiser dans l'interface-administrateur les 280 traits langagiers étiquetés dans Elan pour assurer leur cherchabilité via l'interface-usager ?

b. Comment réduire au minimum l'usage du métalangage linguistique, alors que la finalité de l'interface consiste précisément à présenter des phénomènes linguistiques ?

Dans notre communication, après avoir brièvement présenté le type de documents sonores retenus pour constituer la base de données ainsi que les critères de sélection des traits langagiers du français parlé, nous détaillerons la manière dont nous avons essayé de répondre à ces deux défis, sachant que pour a) la page d'accueil de FLORALE propose les quatre catégories suivantes, qui, en dépit de sa simplicité, autorise l'accès aux 12000 annotations correspondant aux 208 phénomènes annotés à ce jour (sur quatre heures de documents sonores) :

FIG. 2 : Les quatre grandes catégories de la page d'accueil de l'interface-usager de FLORALE

Enfin, si le temps le permet, nous confronterons l'allure et le fonctionnement de notre interface aux divers critères répertoriés par Johnson (2014 : xv) dans ses « Two Best-Known Lists of User-Interface Design Guidelines ».

Références

Anthony, Laurence. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, 30-2, 141-161.

Bond, Z. S. & Garnes, Sara. (1980). Misperceptions of fluent speech. In : Cole, R. A. (Ed.) : Perception and production of fluent speech. Hillsdale : Lawrence Erlbaum, 115-132.

7. En accord avec la réflexion de Anthony (2013 : 158) : « corpus tools are critical to the success of all corpus-based and corpus driven research projects, as well as Data-Driven Learning (DDL) approaches in the classroom ».

Boulton, Alex & Tyne, Henry. (2014). *Des documents authentiques aux corpus. Démarches pour l'apprentissage des langues.* Paris : Didier.

Cuq, Jean-Pierre & Gruca, Isabelle. (2002). *Cours de didactique du français langue étrangère et seconde.* Grenoble : Presses Universitaires de Grenoble.

Cutler, Anne. (2012). *Native Listening : Language Experience and the Recognition of Spoken Words.* Cambridge : MIT Press.

Debaisieux, Jeanne-Marie. (2009). Des documents authentiques oraux aux corpus : un défi pour la didactique du FLE. *Mélanges Crapel*, 31, 35-56.

Debaisieux, Jeanne-Marie & Boulton, Alex. (2007). Alors la question c ' est...? Questions pragmatiques et annotation pédagogique des corpus. *Cahiers de l'AFLS*, 13-2, 31-59.

Detey, Sylvain ; Lyche, Chantal ; Tchobanov, Atanas ; Durand, Jacques & Laks, Bernard. (2009). Ressources phonologiques au service de la didactique de l ' oral : le projet PFC–EF. *Mélanges Crapel*, 31, 223-236.

Gadet, Françoise. (1996). *Le français ordinaire.* Paris : Armand Colin.

Holec, Henri. (1990). Des documents authentiques, pour quoi faire ? *Mélanges Crapel,* 20, 65-74.

Johnson, Jeff. (2014). *Designing with the mind in mind simple guide to understanding user interface design guidelines.* Waltham : Morgan Kaufmann.

Levelt, Willem J. M. (1989). Hochleistung in Millisekunden –Sprechen und Sprache verstehen. *Universitas*, 44-511, 56-68.

Levelt, Willem J. M. (1994). The skill of speaking. In : Bertelson, P., *et al.* (Eds.) : International Perspectives on Psychological Science. Volume I : Leading themes. Hillsdale : Lawrence Erlbaum Associates, 89-104.

Parpette, Chantal. (2008). De la compréhension orale en classe à la réception orale en situation naturelle : une relation à interroger. *Les Cahiers de l ' Acedle*, 5-1, 219-232.

Parpette, Chantal. (2018). Quelle relation entre discours oral naturel et document oral authentique en FLE ? *Action Didactique*, 1, 18-30.

Porcher, Louis. (1995). *Le français langue étrangère : émergence et enseignement d'une discipline.* Paris : Hachette.

Raso, Tommaso & Mello, Heliana. (2014). Introduction. Spoken corpora and linguistic studies : Problems and perspectives. In : Raso, T. & Mello, H. (Eds.) : Spoken corpora and linguistic studies. Amsterdam : John Benjamins, 1-24.

Ravazzolo, Elisa ; Traverso, Véronique ; Jouin, Émilie & Vigner, Gérard. (2015). *Interactions, dialogues, conversations : l'oral en français langue étrangère.* Paris : Hachette.

Vialleton, Élodie & Lewis, Tim. (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.

Worthington, Debra L. & Fitch-Hauser, Margaret E. (2018). *Listening : processes, functions, and competency.* Oxon : Routledge.

Le lexique causatif français et ses équivalents en chinois : Corpus, Méthodologie, Résultats

Ping-Hsueh Chen
LIDILEM, Université Grenoble Alpes
ping-hsueh.chen@ univ-grenoble-alpes.fr

1 Introduction

L'objectif de cette communication sera de comparer les moyens morphosyntaxiques d'expression de la causalité en français et en chinois. Pour ce faire, nous partirons de l'*Échelle de compacité (Scale of compactness)* du typologue australien R.M.W. Dixon (2000). Cette échelle range les mécanismes causatifs dans les langues du plus compact au moins compact, à savoir : les verbes causatifs (ang : *walk, melt* ; fr : *causer, provoquer*) ; les morphèmes causatifs (ang : *lie/lay* ; fr : *simpl**ifier**, modern**iser***) ; le prédicat complexe (*faire* + *Vinf*) et les périphrases causatives moins grammaticalisées (ang : *make somebody cry* ; fr : *forcer qqn à* + *Vinf*). Ce classement des mécanismes causatifs constitue un filtre efficace pour l'étude de la causalité dans les langues (cf. I. Novakova, 2015 : 106-107). Nous l'avons appliqué à l'analyse du fonctionnement des mécanismes causatifs français en comparaison avec le chinois.

2 Corpus et méthodologie

Notre étude contrastive s'appuie sur un corpus parallèle bilingue (français → chinois). Ce dernier est composé de 5 sous-corpus existant sur la plateforme *Sketch Engine*(`https://www.sketchengine.eu/`) : deux corpus de manuels de logiciels (KDE4 et OpenOffice3), un corpus de sous-titres des films (OpenSubtittles2011) et deux corpus de textes institutionnels (UN et MultiUN).

D'un point de vue méthodologique, nous avons d'abord sélectionné des constructions et des verbes causatifs français à partir de travaux antérieurs sur la causalité (cf. A. Jackiewicz, 1998 ; A. Nazarenko, 2000 ; G. Gross et al., 2009 ; S. Diwersy & J. François, 2011 M. Bak Sienkiewicz, 2016). Pour compléter cette liste, nous avons ensuite cherché dans l'ensemble des corpus FR-CH des verbes causatifs en utilisant, par exemple, la grammaire **[lemma=".*iser"& tag="V.*"]** pour la récolte des verbes morphologiques suffixés en *-iser*. Notre liste totalise 136 verbes causatifs et constructions causatives, répartis dans les quatre mécanismes causatifs (cf. la figure 1). Notons ici que, vu la taille importante de notre corpus parallèle bilingue, le seuil de fréquence a été fixé à 1000 occurrences pour chaque mécanisme :

Pour ce qui est du recueil d'occurrences, différentes grammaires sont utilisées pour chaque mécanisme causatif afin de recueillir des données sur la plateforme *Sketch Engine*. Par exemple, la grammaire **[lemma_lc="Verbe en question"& tag="V.*"]** est employée pour extraire les occurrences des verbes lexicaux (1^{er} palier : *causer, provoquer, etc.*) et morphologiques (2^{me} palier : *simpl**ifier***, *modern**iser***, *etc.*). Une autre grammaire est destinée à récupérer des occurrences du prédicat complexe (3^{me} palier), par exemple, *faire* + *sentir* : **[lemma="faire"] [tag="R.*"]***

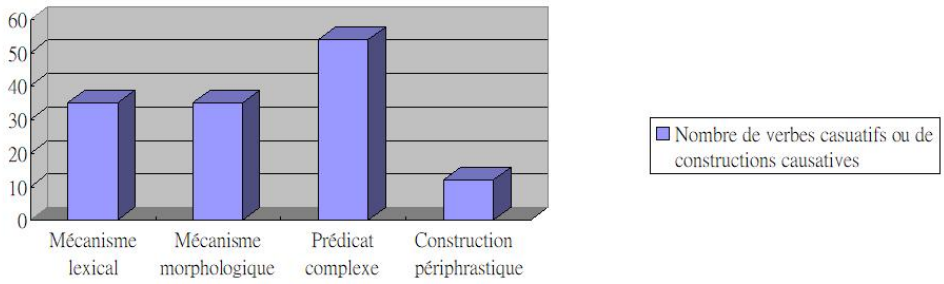


FIG. 1 : Nombre des verbes causatifs ou des constructions causatives étudiés

[lemma="me|te|se|nous|vous"]* [lemma="sentir"]. Pour collecter les périphrases causatives (4^{me} palier), trois grammaires sont utilisées. La première sert à récupérer les occurrences des périphrases causatives du type *contraindre quelqu'un à + Vinf* (avec une préposition) : **[lemma = "me | te | le | la | l' | nous | vous | les"] * [lemma = "avoir"] * [lemma = "contraindre" & tag = "V.*"] [tag = "R.*"] * [tag = "N.*"] * [lemma = "à"][tag = "V.*"].** Deux autres grammaires servent à extraire les occurrences des périphrases causatives du type *permettre à quelqu'un de + Vinf* (avec deux prépositions) :

a. **[lemma = "me | te | lui | nous | vous | leur"][lemma = "avoir"] * [tag = "R.*"]* [lemma = "permettre"& tag = "V.*"][tag = "R.*"]*[lemma = "de"][tag = "V.*"]**

b. **[lemma = "avoir"] * [tag = "R.*"] * [lemma = "permettre"& tag = "V.*"] [tag = "R.*"] * [lemma = "à"] [tag = "N.*"] [lemma = "de"] [tag = "V.*"]**

Par ailleurs, notre corpus parallèle, déjà aligné et syntaxiquement étiqueté, nous facilite l'extraction des équivalents chinois du lexique verbal causatif français. Notons ici que la désambiguïsation sémantique du lexique verbal causatif français s'est effectuée manuellement.

3 Résultats

Le lexique verbal français peut être traduit en chinois par des verbes comme 引起 *yǐnqǐ* (*entraîner*), 造成 *zàochéng* (*causer, provoquer*), 导致 *dǎozhì* (*conduire à*), *etc.* et aussi par des périphrases causatives, telles que 使 *shǐ* (*faire en sorte que*) + *V2 non causatif*, 让 *ràng* (*laisser*) + *V2 non causatif, etc.* Remarquons que les périphrases causatives chinoises sont un des moyens les plus souvent employés pour exprimer la causalité (B. Basciano, 2010, 121-123 ; 2013 : 57).

Les deux verbes causatifs par excellence *provoquer* et *causer* sont traduits en chinois respectivement par un verbe causatif (引起 *yǐnqǐ*, *entraîner*) dans l'exemple (1) et par une périphrase causative moins grammaticalisée (使 *shǐ*, *faire en sorte que* + 陷入 *xiànrù*, *s'enfoncer*) dans l'exemple (2) :

a. Elle **avait provoqué** des remous dans le monde politique [⋯].
(Elle = une lettre de félicitations adressée par le Premier Ministre aux dirigeants des colonies) (*Sketch Engine, MultiUN*)

总理/的/这/封/信/引起/左翼/的/政治/骚动，[…]。(*Sketch Engine, MultiUN*)
”zǒnglǐ / de / zhè / fēng / xìn / yǐnqǐ / zuǒyì / de / zhèngzhì / sāodòng, […].” Litt. Premier Ministre / DE ^{marqueurdepossession} / ce / CL / lettre / **entraîner** / parti de gauche / DE ^{marqueurdepossession} / politique / remous, […]

b. Ça va me **causer** des ennuis. (*Sketch Engine, OpenSubtittles2011*)
你/将/使/我/陷入/麻烦。(*Sketch Engine, OpenSubtittles2011*)
”nǐ / jiāng / shǐ / wǒ / xiànrù / máfán.”
Litt. Tu / jiāng ^{marqueurdefutur} / **faire en sorte que** / je / **s'enfoncer** / ennuis

L'analyse des résultats montre que le français privilégie le prédicat complexe (*faire + Vinf*) pour exprimer la causalité. Alors qu'en chinois, elle est exprimée le plus souvent par des périphrases causatives (*V1 causatif + V2 non causatif*). Suite aux résultats obtenus, nous proposerons un éventail des équivalents fonctionnels chinois des constructions et des verbes causatifs français. Cet éventail permet de mieux appréhender comment le chinois traduit la causalité exprimée par le lexique causatif français.

Références bibliographiques

Bak Sienkiewicz, M. (2016). *Les constructions Verbe causatif + Nom d'émotion. Aspects linguistiques et pistes didactiques* (Thèse de doctorat). Université Grenoble Alpes.

Basciano, B. (2010). *Verbal compounding and causativity in Mandarin Chinese* (Thèse de doctorat). Università di Verona.

———. (2013). Causative light verbs in Mandarin Chinese (and beyond). *Morphology in Toulouse. Selected Proceedings of Décembrettes* 7, *LINCOM Europa*, 57-89.

Dixon, R.M.W. (2000). A typology of causatives : form, syntax and meaning. Dans R.M.W. Dixon & A. Aikhenvald (éds.), *Changing valency. Case studies in transitivity* (p. 30-83). Cambridge : Cambridge University Press.

Diwersy, S., & François, J. (2011). La combinatoire des noms d'affect et des verbes supports de causation en français. Étude de leur attirance au niveau des unités et de leurs classes syntactico-sémantiques. *Revue TRANEL (Travaux neuchâtelois de linguistique)*, 55, 139-161.

Gross, G., Pauna, R. & Valetopoulos, F. (2009). *Sémantique de la cause.* Leuven/Paris : Peeters.

Jackiewicz, A. (1998). *L'expression de la causalité dans les textes Contribution au filtrage sémantique par une méthode informatique d'exploration contextuelle* (Thèse de doctorat). Université de Paris-Sorbonne (Paris IV).

Novakova, I. (2015). *Syntaxe et sémantique des prédicats - Approche contrastive et fonctionnelle.* Éditions universitaires européennes.

Nazarenko, A. (2000). *La cause et son expression en français.* Paris : Ophrys.

Corpus : Sketch Engine : `https://www.sketchengine.eu/`

Erreurs d'apprenants : typologie et annotations

Amalia Todirascu , Marion Cargill et Ioana Buhnila . LiLPa, Université de Strasbourg
todiras@unistra.fr, mcargill@unistra.fr, ibuhnila@unistra.fr

1 Introduction

Nous présentons un travail de recherche visant à annoter les erreurs dans un corpus de productions écrites d'apprenants du FLE. Ce corpus, annoté en niveau CECR (A1-C2), disponible en format électronique, a été créé dans le cadre du projet SimpleApprenant[1]. L'objectif de ce projet est de construire une plateforme d'aide à l'apprentissage des langues, qui intègre outils et ressources de Traitement Automatique des Langues. Ainsi, la plateforme analyse des productions écrites des apprenants et des stratégies de correction et de remédiation des erreurs. Pour ce faire, nous avons étudié et annoté manuellement les erreurs d'un corpus d'apprenants du FLE.

L'objectif de ce travail d'annotation des erreurs d'apprenants est double : d'une part, nous proposons un corpus annoté en niveau CECR (A1-C2), en mettant en évidence les erreurs lexicales, syntaxiques et stylistiques. Ce corpus permet d'identifier les types d'erreurs les plus fréquentes selon les locuteurs natifs de plusieurs langues (polonais et grec). D'autre part, le corpus est analysé dans l'objectif de créer des règles de correction et de transformation des productions des apprenants, en proposant un retour immédiat et une explication des erreurs.

En effet, les erreurs montrent des difficultés liées à la maîtrise de la L2 et à l'influence du L1 du locuteur. Le repérage des erreurs fréquentes permet la mise en place des stratégies pédagogiques pour les éviter (Granger et al, 2015). Pour ce faire, les corpus d'apprenants représentent des ressources très intéressantes, en particulier si les erreurs sont annotées (Boulton et Tyne, 2014), (De Cock et Tyne, 2014). Plusieurs corpus d'apprenants annotés en erreurs sont disponibles : ICLE (Granger et al., 2002) pour l'anglais, FRIDA (Granger, 2007) pour le français.

L'annotation d'erreurs est une tâche difficile aussi bien pour les annotateurs humains et pour les outils d'annotation automatique. Certains corpus (FRIDA, Granger, 2007) sont annotés manuellement pour mettre en évidence des erreurs d'apprenants et leur typologie (Dagneaux et al., 1998, 2008). Les erreurs sont classées par domaine (lexical, syntaxique, stylistique) et caractérisées par type d'opération réalisée sur le texte (ajout, changement, suppression, répétition). Pour la correction automatique, les systèmes proposant une annotation automatique des erreurs pour l'anglais (Rozovskaya and Roth, 2010), (Gaillat, 2013) ou pour l'allemand (Kempfert, et Köhn, 2018) se concentrent sur des phénomènes spécifiques et s'appuient sur des corpus annotés en parties de discours ou en syntaxe (Bryant et al., 2017). Ces projets adoptent des typologies d'erreurs simplifiées adaptés à la tâche de découverte des erreurs. Dans le même contexte, notre étude vise à identifier des règles de correction automatique d'erreurs, par conséquent nous avons établie une typologie à partir de nos données, avec quelques différences par rapport à (Dagneaux et al, 2008).

1. https://simpleapprenant.huma-num.fr/SimplifyYourFrench/accueil

2 Corpus et méthodologie

2.1 Corpus

Notre corpus est constitué à l'aide de l'Université d'Opole (Pologne) et de l'Université de Chypre. Ces universités proposent des cours de FLE, pour les futurs enseignants, qui sont des locuteurs natifs de polonais ou de grec. Sur chaque site, nous avons recueilli des productions écrites numériques anonymisées, réalisées en classe ou à la maison. Le corpus est constitué de plusieurs niveaux provenant des deux sites :

	A1		A2		B1		B2-C1	
	Tokens	Erreurs	Tokens	Erreurs	Tokens	Erreurs	Tokens	Erreurs
Pologne	13715	1164	15416	1312	4527	365	11958	1304
Chypre	7749	651	8505	781	5755	472	17165	1288

TAB. 1 : La taille de corpus (en nombre de tokens) et le nombre d'erreurs

3 Méthodologie

Pour créer notre corpus, nous avons rassemblé des textes écrits d'apprenants du FLE, à l'aide de SimpleApprenant. Nous avons analysé manuellement les erreurs d'un échantillon de textes provenant de deux sites différents, représentatif des niveaux A1-C1. Ainsi, nous proposons une typologie d'erreurs et nous la comparons avec les typologies existantes (Dagneaux et al., 1998, 2008). Un guide d'annotation illustrant les types d'erreurs repérées est utilisé pour l'annotation des erreurs par deux annotateurs et pour adjudication par un expert.

Dans un premier temps, nous avons analysé les erreurs sur un échantillon de 12 textes pour chaque site (4 textes par niveau). Nous avons classé les erreurs en plusieurs domaines, divisés en plusieurs catégories (Dagneaux et al., 1998, 2008) : erreurs d'orthographe (lettre ajoutée ou manquante, manque d'accents), erreurs lexicales (choix erroné d'un mot ou de la catégorie lexicale, ajout d'un mot, suppression d'un mot), erreurs grammaticales (absence d'accord entre le déterminant et le nom, absence de déterminant, confusion entre les temps et les mode pour le verbe), erreurs syntaxiques (ordre des mots, sujets ou objets manquants, confusion du pronom, etc.), erreurs de style. La proportion des erreurs évolue avec le niveau : le niveau A1/A2 est riche en erreurs d'orthographe alors que les niveaux B1/B2/C1 sont plus riches en erreurs syntaxiques ou stylistiques. Nous avons fait des choix différents de classification par rapport à (Dagneaux et al, 1998, 2008) : les erreurs de registre et de style sont classées dans le même domaine, la ponctuation est une catégorie du domaine syntaxique et les opérations sont inclues dans les catégories.

4 Résultats

Sur la base de ces observations et de la typologie, nous avons constitué un guide d'annotation des erreurs. Le guide regroupe des exemples illustrant chaque type d'erreur et sa délimitation. Deux annotateurs ont travaillé sur un corpus constitué de 52 textes suivant le guide d'annotation. Ainsi, ils ont annoté le plus petit fragment qui contient l'erreur et la première erreur du même type dans la phrase. Les erreurs sont représentés dans un fichier csv, indiquant le domaine et le type d'erreur et une solution de remédiation. Ce fichier est utilisé par un script Perl pour générer les fichiers annotés avec l'erreur et la correction (en format XML). Nous avons comparé le travail des deux annotateurs et nous avons analysé les cas de désaccord. L'accord inter-annotateur a été calculé sur 12 textes, permettant de revoir le guide. 207 erreurs sont identifiées dans ces textes par les deux annotateurs. Seulement 124 erreurs (59,90 %) sont reconnues ayant le même domaine et la même catégorie. 83 erreurs (40,10 %) sont reconnues mais classées différemment par les deux annotateurs. Parmi les erreurs identifiées conjointement par les deux annotateurs, on constate que :

- 10,44 % représentent les erreurs d'orthographe (manque d'accents) ;
- 17,91 % sont des erreurs grammaticales (confusion de préposition, adverbe ou conjonction) ;
- 19,41 % représentent des erreurs grammaticales (erreurs de conjugaison) ;
- 14,92 % sont des erreurs grammaticales concernant l'utilisation des articles (mauvais nombre, absence d'article, substitution d'un article) ;
- 25,37 % sont des erreurs de syntaxe (toutes erreurs confondues, plus de la moitié concernent l'ordre de mots et la ponctuation) ;
- 4,47 % d'erreurs visant l'accord en nombre ou en genre ;
- les confusions de parties de discours représentent 7,40 % des erreurs.

En ce qui concerne les divergences, la plupart des erreurs sont des différences de délimitation (44 %) mais aussi de classement. Les divergences de classement les plus fréquentes sont constatés entre le niveau lexical et le niveau de l'orthographe ou entre la syntaxe et le style.

Le travail d'annotation sera étendu à un corpus d'apprenants FLE, locuteurs espagnols, pour compléter le corpus annoté. Nous comparerons les types d'erreurs et leur distribution entre les différents niveaux CECR entre les trois sous-corpus (polonais, grec, espagnol).

Références bibliographiques

Bryant, C., Felice, M, Briscoe, T. (2017). Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Volume 1 : LongPapers), pages 793–805, Vancouver, Canada.

Boulton, A. & Tyne, H. (2014). *Des documents authentiques aux corpus : démarches pour l'apprentissage des langues.* Paris : Didier.

Dagneaux, E., Denness, S. & Granger, S. (1998). Computer-aided error analysis. *System*, vol. 26. 163-174.

Dagneaux, E., Denness, S., Granger, S., Meunier, F., Neff, J. & Thewissen, J. (2008). Error tagging manual, version 1.3. Centre for English Corpus Linguistics. Louvain-la-Neuve : Université Catholique de Louvain.

De Cock, S., Tyne H. (2014). Corpus d ' apprenants et acquisition des langues, *Recherches en Didactique des Langues et Cultures : les Cahiers de l' ACEDLE*, 11(1), pp.137-168.

Gaillat, T. (2013). Annotation automatique d'un corpus d'apprenants d'anglais avec un jeu d'étiquettes modifié du Penn Treebank. *Actes de 20e conférence sur le TALN*, 271-284.

Granger, S., Dagneaux E., Meunier. F. (2002). *International Corpus of Learner English*

Granger, S. (2007) Corpus d'apprenants, annotation d'erreurs et ALAO : une synergie prometteuse. *Cahiers de Lexicologie*, Vol. 91, no. 2, p. 117-132.

Granger S., Gilquin G., Meunier F. (2015). *The Cambridge Handbook of Learner Corpus Research.* Cambridge : Cambridge University Press

Rozovskaya A., Roth, D. (2010) Annotating ESL Errors : Challenges and Rewards, *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 28–36, Los Angeles, California

Kempfert, I., Köhn, C. (2018). An automatic error tagger for German. *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning at SLTC 2018 (NLP4CALL 2018)*. Linköping Electronic Conference Proceedings 152 : 32–40

Session 2.A.
Analyse de corpus oral : enseignement de la L2

Des liaisons et des corpus : apports d'une étude sur le changement linguistique en temps réel

Céline Dugua [1], Jennifer Ganaye [1], Flora Badin [1] et Olivier Baude [2]
[1] Laboratoire Ligérien de Linguistique, UMR7270, Université d'Orléans
[2] MoDyCo, UMR7114, Université Paris Nanterre
celine.dugua@univ-orleans.fr, jennifer.ganaye@univ-orleans.fr, flora.badin@univ-orleans.fr,
olivier.baude@parisnanterre.fr

1 Introduction

La liaison en français est un phénomène largement documenté (Chevrot, Fayol & Laks, 2005 ; Delattre, 1956 ; Schane, 1967), son caractère variable en fait par ailleurs un objet phare des études en sociolinguistique (Armstrong, 2001 ; Encrevé, 1988 ; Gadet, 1989). L'apport des corpus pour observer l'usage des liaisons a permis de revisiter les classements des ouvrages de références classiques, notamment sur la question des contextes de liaison « obligatoires » et « facultatifs » (Baude & Dugua, 2015). Parmi les corpus à partir desquels la liaison a été étudiée, citons en sept qui permettent de mettre en évidence la diversité des axes d'observation, tant au niveau des situations que des locuteurs :

- Ågren (1973) : les liaisons dans des conversations radiophoniques ;
- Encrevé (1988) : les liaisons chez les hommes politiques et notamment les liaisons sans enchainement ;
- De Jong (1988, 1994) : les liaisons dans 45 entrevues du corpus d'Orléans et dans des données du corpus de Tours ;
- Durand, Laks & Lyche (2003), Durand & Lyche (2008) : mise en place du corpus PFC spécialement dédié à l'étude de l'usage de la liaison dans différents points d'enquêtes et sous différentes modalités langagières (entretiens, lecture d'un texte, lecture de mots) ;
- Chabanal (2003) : corpus de deux enfants âgés de 40 à 50 mois issus de milieux sociaux contrastés ;
- Nardy (2008) : corpus recueilli dans une classe d'école maternelle avec prise en compte des interactions au sein du réseau de pairs ;
- Liégeois (2014) : corpus dense d'interactions parents-enfants dans leur environnement familial.

Les analyses sur corpus sont donc aujourd'hui nombreuses et variées mais il reste de nombreuses données à explorer. Ainsi le corpus des Enquêtes SocioLinguistiques à Orléans offre l'opportunité d'une analyse diachronique inédite.

2 Corpus et méthodologie

2.1 Corpus

Pour comprendre ce que peut apporter une étude sur le corpus ESLO, il est nécessaire de présenter quelques-unes de ses caractéristiques. Le corpus des Enquêtes sociolinguistiques à Orléans regroupe deux enquêtes réalisées à 40 ans d'intervalle : ESLO1 dans les années 1968-74 et ESLO2 depuis 2006. ESLO1 a été constitué par une équipe d'enseignants anglais qui cherchait à observer et capter la dynamique des pratiques linguistiques partagées par les habitants d'une cité afin de construire le « portrait sonore d'une ville ». La démarche était dès le départ ancrée dans le champ de la sociolinguistique et de la prise en compte de la variation (Baude & Dugua, 2016).

> Selon nous une recherche sociolinguistique impliquait une étude de la langue dans sa diversité plutôt que comme un tout homogène et figé. En effet, même si on étudie un état de langue à un moment précis de l'histoire, il n'empêche qu'il offre une variété à plusieurs niveaux : différences entre les générations, différences dialectales entre communautés, différences entre les milieux sociaux, différences liées aux conditions de production du discours. (Blanc & Biggs, 1971 :16)

La méthode de constitution d'ESLO2 prend en compte l'expérience d'ESLO1 et également l'évolution des cadres théorique et méthodologique. L'objectif étant de mettre à disposition un grand corpus variationniste, (1) mutualisant les données d'ESLO1 et d'ESLO2, (2) intégrant différentes situations de recueil (différents modules) dans une architecture prenant en compte le degré de formalité de la situation et la classification des locuteurs (Baude & Dugua, 2011) (entretiens, paroles publiques, conversations lors de repas, réunions de travail etc. mais aussi un sous-corpus composé d'enregistrements de mêmes locuteurs à quarante années de distance) et (3) rendant toujours disponibles pour le chercheur les informations sur les participants et sur les situations.

Les transcriptions de ce corpus respectent des conventions « neutres » qui constituent l'annotation de premier niveau. Pour chaque enregistrement, trois versions de transcription sont réalisées et conservées, aucune n'intègre d'annotations plus fines que l'orthographe et nos conventions (voir Guides du transcripteur : `http://eslo.huma-num.fr/index.php/pagetranscription`).

La liaison a déjà fait l'objet de travaux sur une partie de ce corpus (le corpus d'Orléans, devenu ESLO1) dans une étude majeure consacrée à la sociophonologie de la liaison (De Jong, 1988, 1994). L'auteur observe à la fois les usages de la liaison dans un grand nombre de contextes lexicaux, mais aussi l'influence de facteurs sociaux (âge, sexe, CSP) sur la réalisation de cette variable. Il s'agissait d'un travail novateur sur la liaison dans le sens où l'accent était mis non pas sur des catégories morpho-syntaxiques mais sur des formes lexicales. Par exemple, dans la catégorie des adverbes monosyllabiques (classés comme obligatoire chez Delattre, 1966) avec une moyenne de 92% de liaisons réalisées dans son corpus, De Jong a pu montrer une variation importante à l'intérieur de cette catégorie : d'un taux équivalent à 99,4% pour l'adverbe « très » à un taux de 54,5% pour « mieux ». Cette façon d'observer l'usage de la liaison nous intéresse

puisqu'elle pose, au centre du traitement, des formes concrètes plus que des catégories, comme le suggèrent les théories basées sur l'usage (Tomasello, 2003).

L'accès au corpus d'origine avec des outils d'exploitation nous permettra de revenir sur l'étude de De Jong. Il ne s'agira pas de vérifier de façon stricte l'analyse des données réalisée à l'époque, mais plutôt de les revisiter avec un nouvel éclairage tant théorique que méthodologique. La mise en œuvre d'ESLO2 permettra une étude sur les variations diachroniques, diastratiques et diaphasiques à partir d'un très grand corpus (Baude & Dugua, 2011).

2.2 Méthodologie

ESLO1 et ESLO2 ont permis de constituer un sous-corpus d'ESLO pour observer le changement en temps réel : le module Diachronie. Dans ce module sept locuteurs ont été enregistrés à deux reprises : une première fois dans les années 1970 dans le cadre du projet ESLO1, puis une deuxième fois dans les années 2007 par Vaslin-Chesneau (2008) pour le compte du projet ESLO2. Il s'agit donc d'un sous-corpus original dans lequel nous pourrons observer d'éventuels effets de variation diachronique. Par ailleurs, pour chaque locuteur, nous disposons d'informations et de données socio-démographiques pour chacune des deux périodes. Ce point est important et nous pourrons le détailler lors de notre présentation, car en 40 ans, les caractéristiques notamment sur la profession, ont pu évoluer de manière significative. Pour chacune des enquêtes, les locuteurs ont participé à des entretiens semi-directifs autour de thématiques sur la famille, le travail, les activités dans la ville et les pratiques sociales et culturelles. La trame d'entretien utilisée pour ESLO2 a été élaborée pour permettre une comparaison précise ; elle calque en l'ajustant le questionnaire d'ESLO1. Ce sous-corpus comprend 14 enregistrements d'une durée totale d'environ 17 heures (10 heures pour ESLO1 et 7 heures pour ESLO2).

En partant des transcriptions de ces 14 enregistrements, nous avons mis en place une procédure permettant de repérer automatiquement tous les contextes de liaison potentiels, en s'appuyant au départ sur une simple définition graphique : un mot (mot1) qui se termine par une consonne suivi d'un mot (mot2) qui commence par voyelle ou « h ». Nous obtenions un tableau nous donnant pour chaque contexte mot1-mot2, le nom du fichier sonore, le code locuteur, le time code, le contexte élargi, la catégorie morpho-syntaxique des deux mots (Dugua et al. 2017). Ce tableau nous sert ainsi de guide pour retrouver chaque contexte repéré et pour écouter si la liaison a été réalisée ou non, et sous quelle forme (enchainement ou non enchainement, pataquès).

3 Résultats

Les données sont en cours de traitement. L'originalité de notre approche sera de mener des analyses à différents niveaux : (1) d'un point de vue quantitatif sur un corpus oral d'une durée de 17 heures pour lequel on peut estimer le nombre de contextes de liaison à environ 14.000 ; (2) mais aussi avec une approche variationniste grâce aux informations disponibles sur les participants et sur leurs trajectoires de vie durant les 40 ans qui séparent la première enquête de la seconde ; (3) avec une approche en temps réel de l'usage de la liaison, afin de confirmer les premiers résultats (Baude et Dugua, 2015) qui tendaient à montrer que globalement l'usage de la

liaison n'est pas en baisse, mais que la variation individuelle suggéraient un changement linguistique en temps apparent ; (4) au regard également des contextes morpho-syntaxiques et lexicaux qui permettront d'affiner et de discuter les classements existants et de mettre en perspectives nos résultats d'une étude lexicale avec les travaux de De Jong.

Références bibliographiques

Ågren, J. (1973). *Étude sur quelques liaisons facultatives dans le français de conversation radiophonique : fréquences et facteurs.* Uppsala : Acta Universitatis Upsaliensis.

Armstrong, N. (2001). *Social and stylistic variation in spoken French : a comparative approach.* Amsterdam, Philadelphia : John Benjamins.

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

Baude, O. & Dugua, C. (2015). Usage de la liaison dans le corpus des ESLOs : vers de nouveaux (z) ouvrages de référence ? In Dostie, Hedermann (Eds.). *La dia-variation en français actuel.* Bern : Peter Lang, p. 349-371.

Baude, O. & Dugua, C. (2016). Les ESLO, du portrait sonore au paysage digital, *Corpus*, 15, 29-56.

Blanc, M. et Biggs, P. (1971). L' enquête socio-linguistique sur le français parlé à Orléans. *Le Français dans le Monde*, 85, 16–25.

Chabanal, D. (2003). *Un aspect de l'acquisition du français oral : la variation socio-phonétique chez l'enfant francophone.* Thèse de doctorat en Sciences du Langage, Université Paul-Valéry, Montpellier.

Chevrot, J.-P., Fayol, M. & Laks, B. (2005). La liaison : de la phonologie à la cognition. *Langages*, 158, 3-7.

De Jong, D. (1988). *Sociolinguistic Aspects of French Liaison,* Free University Amsterdam. Unpublished Ph.D. Thesis.

De Jong, D. (1994). La sociophonologie de la liaison orléanaise. In Lyche, C. (Ed.), *French Generative Phonology : Retrospective and Perspectives* (pp.95-129). Salford : ESRI.

Delattre, P. (1966). *Studies in french and comparative phonetics : selected papers in French and English.* The Hague, London, Paris : Mouton & Co.

Dugua, C. & Baude, O. (2017). La liaison à Orléans, corpus et changement linguistique : une première étude exploratoire, *Journal of French Language Studies,* 27, 41-54.

Durand, J., B. Laks & C. Lyche (2003) Le projet Phonologie du français contemporain. *La Tribune Internationale des Langues Vivantes*, 33, 3-9.

Durand, J. & Lyche, C. (2008). French Liaison in the Light of Corpus Data. *Journal of French Language Studies*, 18-1, 33-66.

Encrevé, P. (1988). *La liaison avec et sans enchaînement, phonologie tridimensionnelle et usage du français.* Paris : Edition du Seuil.

Gadet, F. (1989). *Le Français ordinaire.* Paris : Armand Colin.

Liégeois, L. (2014). *Usage des variables phonologiques dans un corpus d'interactions naturelles parents-enfant : impact du bain linguistique et dispositifs cognitifs d'apprentissage.* Thèse de doctorat, Université Blaise Pascal, Clermont-Ferrand.

Nardy, A. (2008). *Acquisition des dialectes sociaux et des usages académiques entre 2 et 6 ans : facteurs socio-démographiques et influence du groupe de pairs,* Thèse de doctorat, Grenoble, Université Stendhal. `http://tel.archives-ouvertes.fr/tel-00466276/fr/`

Schane, S. A. (1967). L'élision et la liaison en français. *Langages*, 8, 37-59.

Tomasello, M. (2003). *Constructing a language : a usage-based theory of language acquisition.* Cambridge, Massachussetts : Harvard University Press.

Vaslin-Chesneau, A. (2008). *Analyse diachronique de la variation sociolinguistique à partir de deux corpus orléanais.* Thèse de doctorat, Université d'Orléans.

Site ESLO : `http://eslo.huma-num.fr/`

Corpus oral longitudinal d'apprenants de Français L2 en immersion. Enjeux méthodologiques d'annotation et d'analyse de la production orale.

Minerva Rojas Madrazo
LLSETI EA 3706 Université Savoie Mont Blanc
rojasmam@univ-smb.fr

1 Introduction

L'étude de corpus oraux longitudinaux contribue de manière importante dans la recherche en acquisition des langues, permettant d'observer de manière détaillée le processus d'apprentissage (Myles, 2005 ; DeCock & Tyné, 2014). Suivant des principes théoriques et méthodologiques de la Théorie de la Complexité et des Systèmes Dynamiques (Verspoor, de Bot & Lowie, 2011 ; Lowie, 2017) l'adoption d'une méthodologie longitudinale nous permet de tracer la trajectoire individuelle, la variabilité ainsi que les phénomènes de stabilisation et de réorganisation de l'apprentissage (Larsen-Freeman & Cameron, 2008). À la différence des études transversales, les études longitudinales ne cherchent pas la généralisation des résultats, mais la description exhaustive d'une variété de facteurs qui montrent la complexité du développement de la L2.

2 Corpus et méthodologie

2.1 Corpus

Nous avons créé un corpus longitudinal qui recueille les productions orales d'un groupe d'apprenants de FL2 en immersion (n=12), tous inscrits dans la même licence à l'Université Savoie Mont Blanc. Les participants ont effectué huit tâches au total, divisées en quatre collectes de données entre 2015 et 2017 ; dans chacune des collectes, ils ont accompli une tâche narrative individuelle (*retell story*) et une autre tâche en interaction. Les données orales constituent la base du corpus, créé à l'aide du logiciel EXAMARaLDA (Schmidt, 2004) permettant une transcription et annotation minutieuses, compatibles avec les normes de transcription et les formats CLAN (MacWhinney, 2000). Donc, pour rendre compte de la complexité de l'évolution de la production orale, nous avons focalisé notre étude sur l'analyse longitudinal de trois variables : la fluidité énonciative ou productive, les stratégies de communication et le lexique. Ces variables constituent, en même temps, des constructions théoriques complexes, donc leur étude groupe une série d'unités d'analyse recouvrant leur caractère multidimensionnel.

2.2 Méthodologie

L'étude de la fluidité part de l'annotation de la durée et le nombre de pauses et de disfluences afin de calculer des mesures quantitatives, telles que la vitesse d'articulation, la longueur moyenne des segments (*MLR*) et des énoncés (*MLU*), la durée moyenne de pauses et la proportion de pauses et de disfluences pour 100 mots ; les résultats de ces mesures nous fournissent des indices sur le processus d'automatisation de la L2 (Hilton, 2009 ; Kormos, 2006). D'autre part,

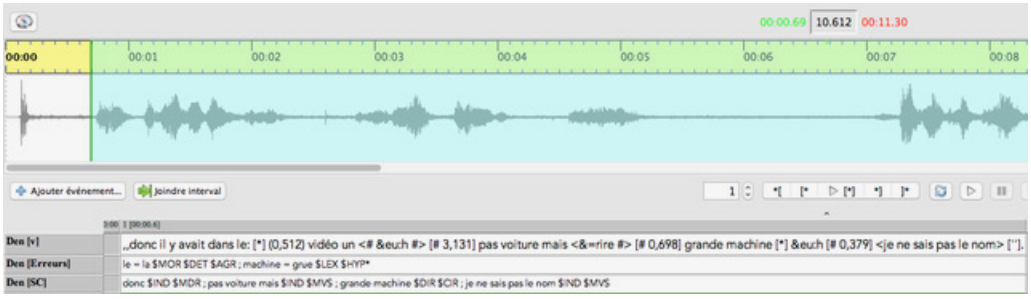


FIG. 1 : Détail de transcription

les stratégies de communication, susceptibles d'être accompagnées de phénomènes de disfluence (Faerch & Kasper, 1983 ; Poulisse, 1993), constituent des ressources utilisées par les locuteurs pour garder le flot de la production et sont associées à des difficultés ou vides linguistiques, à la négociation du sens ou à la gestion du temps lors du traitement langagier (Celce-Murcia, Dörnyei & Thurrell, 1995) ; pour leur analyse, nous avons adapté et créé un code d'annotation à partir de la taxonomie établie par Dörnyei et Scott (1997) qui intègre les trois perspectives mentionnées ; donc leur annotation sert à calculer une série de mesures quantitatives afin d'analyser la distribution et la proportion de stratégies dans la production orale. Enfin, nous avons analysé le profil lexical des participants en calculant, d'une part, la proportion d'erreurs pour 100 mots (Boulté & Housen, 2012) à l'aide de l'annotation d'erreurs phonologiques, lexicales, morphologiques et syntaxiques ; d'autre part, nous avons calculé l'indice D de diversité lexicale (McKee, Malvern & Richards, 2000) et la longueur moyenne des mots (Granfeldt, 2006) à l'aide du logiciel CLAN. Dans le détail de la transcription (1), on voit les paramètres de fluidité annotés dans la première piste de transcription, l'annotation des erreurs est effectuée dans la deuxième piste, et l'annotation des stratégies de communication (SC) dans la troisième piste. Bref, l'annotation détaillée de ces paramètres sert à effectuer la recherche de concordances dans le corpus, et, extraire les données nécessaires pour calculer les unités d'analyse des variables visées et leurs corrélations avec le temps.

Afin de trianguler l'analyse quantitative de la production orale, nous avons également recueilli des données complémentaires (linguistiques et extralinguistiques), inclues comme métadonnées dans le corpus ; elles portent sur des informations personnelles des apprenants (L1, pays d'origine, âge, lieu d'habitation), le contact quotidien avec la langue française, les années d'expérience en L2, le niveau en L2 au début de l'étude, et la motivation d'apprendre la L2. Étant donné que les apprenants suivis ont des profils hétérogènes (dont huit pays d'origine et huit L1, ainsi que des expériences préalables différentes en FL2), les informations complémentaires nous aident à tracer leurs conditions initiales et à interpréter leurs trajectoires individuelles (Verspoor, 2015). De cette manière, les informations portant sur les conditions initiales constituent un point important de l'analyse qualitative et aident à mieux comprendre les résultats quantitatifs et les changements et variations survenus dans le développement de la production orale en L2.

Enfin, même si le but de notre étude de corpus vise l'analyse longitudinal des apprenants, nous avons également menée une analyse contrastive suite à la constitution d'un corpus com-

parable de natifs (n=14) effectuant les mêmes tâches que les apprenants ; l'analyse contrastive nous permet de décrire si la production des apprenants se rapproche des productions des natifs du début à la fin de l'étude, mais, elle constitue davantage une pièce importante de la méthodologie d'identification des stratégies de communication nous appuyant sur la Théorie Basée sur l'Usage (Smiskova-Gustafsson, 2013).

En somme, dans notre communication nous présenterons les enjeux méthodologiques autour de la constitution de notre corpus et du protocole d'analyse de la production orale en L2 depuis une perspective dynamique et complexe. Nous ferons le point sur l'annotation et les unités d'analyse étudiées, et, dans quelle mesure les résultats issus des analyses reflètent l'évolution de la production orale en FL2.

Références bibliographiques

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. Dans House, A., Kuiken, F. & Vedder, I. (dir.), *Dimensions of L2 performance and Proficiency. Complexity, Accuracy and Fluency in SLA,* pp. 21-46. John Benjamins. Amsterdam/Philadelphia.

Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995). Communicative competence : A pedagogically motivated model with content specifications. *Issues in Applied linguistics*, 6 (2), pp. 5-35.

De Cock, S. et Tyne, H. (2014) Corpus d'apprenants et acquisition des langues. *Recherches en Didactique des Langues et Cultures : les Cahiers de l'acedle, Recherches en Didactique des Langues et des Cultures, 11* (1), pp. 137-168.

Dörnyei, Z. & Scott, M. L. (1997). Communication Strategies in a Second Language : Definitions and Taxonomies. *Language Learning. 47* (1), pp. 173–210.

Faerch, C., & Kasper, G. (1983). On identifying communication strategies in interlanguage production. Dans Faerch, C., & Kasper, G. (dir.) *Strategies in interlanguage communication,* pp. 210-238. Longman. Londres.

Granfeldt, J. (2006). Évaluation du niveau lexical et grammatical à l'écrit en français langue étrangère : l'apport des analyses automatiques. *Revue française de linguistique appliquée 1*, Vol. XI, p. 103-117.

Hilton, H. (2009). *Systèmes émergents : acquisition, traitement et didactique des langues.* Habilitation à Diriger de Recherches. Université Lumière-Lyon II.

Kormos, J. (2006). *Speech production and second language acquisition.* Routledge. New York/Londres.

Larsen-Freeman, D., & Cameron, L. (2008). Research Methodology on Language Development from a Complex Systems Perspective. *Modern Language Journal, 92* (2), pp. 200-213.

Lowie, W. (2017) Lost in state space ? Methodological considerations in Complex Dynamic Theory approaches to second language development research. Dans Ortega, L., & Han, Z. (dir.), *Complexity theory and language development : In celebration of Diane Larsen-Freeman*, pp 123-141. John Benjamins. Amsterdam.

MacWhinney, B. (2000). *The CHILDES Project : Tools for analyzing Talk.* 3rd Édition. Mahwanah, NJ : Lawrence Erlbaum Associates.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and linguistic computing, 15* (3), pp. 323-338.

Myles, F. (2005) Interlanguage corpora and second language acquisition research. *Second Language Research, SAGE Publications, 21* (4), pp. 373-391.

Poulisse, N. (1993). A theoretical account of lexical communication strategies. Dans Schreuder, R. et Weltens, B. (dir.) *The bilingual lexicon,* pp. 157-189. John Benjamins. Amsterdam.

Schmidt, T. (2004). Transcribing and annotating spoken language with EXMARaLDA. *Proceedings of the LREC-Workshop on XML based richly annotated corpora*, Lisbon 2004. En ligne : `https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/2317`.

Smiskova-Gustafsson, H. (2013). *Chunks in L2 development : a usage-based perspective.* (Thèse de doctorat.) Université de Groningen. En ligne : `https://www.rug.nl/research/portal/files/14408764/H_Gustafsson_Dissertation.pdf`

Verspoor, M. (2015). Initial conditions. Dans Dörnyei, Z., MacIntyre, P. D., Henry, A. (dir.) *Motivational dynamics in language learning,* pp. 38-46. Multilingual Matters.

Verspoor, M., De Bot, K., & Lowie, W. (2011). *A dynamic approach to second language development : Methods and techniques* (Vol. 29). John Benjamins Publishing. Amsterdam.

Exploitation de corpus oraux et multimodaux pour apprendre ou enseigner à interagir en français langue étrangère

Virginie André
ATILF, Université de Lorraine
Virginie.Andre@univ-lorraine.fr

1 Introduction

Cette étude s'inscrit à l'interface entre la linguistique et la didactique des langues. Plus précisément, elle cherche à exposer les nécessaires liens entre les deux disciplines en montrant de quelles façons des corpus oraux et multimodaux peuvent être exploités à des fins didactiques. Dans cette perspective, l'exploitation des corpus pour enseigner et apprendre une langue prend la forme d'une analyse sociolinguistique outillée afin de dégager des règles de fonctionnement sur le principe du *data-driven learning* (Johns 1991, Aston 2001), traduit en français par Apprentissage Sur Corpus (ASC) par Boulton et Tyne (2014). L'originalité de notre étude réside dans les corpus exploités. Jusqu'à présent le *data-driven learning* a été expérimenté, souvent avec des résultats très positifs, quasiment uniquement en situation d'enseignement et d'apprentissage de la langue anglaise écrite (Timmis 2015 ; Boulton, Cobb 2017). Nous proposons d'utiliser l'apprentissage sur corpus pour enseigner et apprendre à interagir en français langue étrangère (FLE).

Les recherches actuelles en didactique des langues s'accordent pour soutenir que l'exposition à la langue cible pour des apprenants est indispensable. Depuis les années 1970 et les travaux qui prônent l'utilisation des documents authentiques (Duda, Esch, Laurens 1972 ; Abé, Carton, Cembalo, Régent, 1979), de nouvelles formes d'exposition sont apparues notamment avec les progrès technologiques et l'accès au numérique. Nous nous intéresserons ici à une nouvelle méthodologie qui consiste à exploiter des corpus à des fins didactiques (André 2018). En d'autres termes, nous tenterons de montrer de quelles façons l'ASC peut être utilisé à des fins d'enseignement et d'apprentissage des interactions verbales orales.

2 Corpus et méthodologie

Les enseignants de FLE se plaignent fréquemment du manque de ressources orales authentiques représentatives de situations de communication ordinaires ou répondant aux objectifs des apprenants. Pour pallier ce manque, nous proposons d'utiliser notamment des données recueillies et traitées par des linguistiques. Ces données sont constituées en corpus, défini en linguistique notamment par Sinclair (1996 : 4) par « a collection of pieces of language that are selected and ordered according to explicit criteria in order to be used as a sample of language ». C'est donc à cet échantillon de langue que nous proposons d'exposer les apprenants.

Parmi les corpus oraux et multimodaux pouvant être utilisés à des fins didactiques, nous pensons notamment à TCOF (`http://www.cnrtl.fr/corpus/tcof/`) constitué à Nancy, à CLAPI (`http://clapi.ish-lyon.cnrs.fr/`) constitué à Lyon, à ESLO (`http://eslo.huma-num.`

fr/) constitué à Orléans, à OFROM constitué à Neuchâtel en Suisse (http://www11.unine.ch/) ou encore au projet PFC (https://www.projet-pfc.net/). La plateforme ORTOLANG présente également une liste de ce type de corpus (https://www.ortolang.fr/). Ces corpus de français parlé ont été recueillis et traités pour accueillir des analyses (socio)linguistiques et (socio)interactionnelles. Néanmoins, ils présentent des ressources précieuses pour l'enseignement et l'apprentissage du FLE dans la mesure où ils sont composés d'enregistrements (son et/ou vidéos) de situations de communications authentiques, ressources rares et recherchées en didactique.

Toutefois, il ne suffit pas d'avoir accès à des données langagières pour permettre l'apprentissage de la langue. La linguistique de corpus apporte de nouvelles descriptions, de nouvelles méthodologies, de nouvelles connaissances sur la langue qui méritent toutes d'être exploitées en didactique des langues (Conrad 2000). L'utilisation de corpus de données authentiques à des fins didactiques nécessite de repenser la façon d'aborder les documents supports en séance de formation. Tout d'abord, pour les enseignants, il s'agit de savoir observer et analyser ces données afin de rendre possible l'appropriation de la langue. La didactique des langues bénéfice de l'avancée des progrès de la linguistique qui tente de se munir de corpus visant une représentativité du français parlé en interaction ainsi que sa description. La sociolinguistique des interactions verbales (André 2015) propose une description des pratiques langagières qui permettent d'accomplir des activités interactionnelles dans des situations de communication particulières. Plus précisément, cette approche cherche à appréhender les pratiques langagières, en analysant leurs spécificités linguistiques, interactionnelles, pragmatiques et contextuelles. Les analyses proposées, à partir de corpus d'interactions authentiques, sont indiscutablement utiles, voire indispensables, à la didactique des langues. Si elles sont mises à la disposition des enseignants et des apprenants, elles permettent d'apprendre et d'enseigner à interagir de façon appropriée.

En outre, l'ASC s'appuie sur la possibilité pour un apprenant d'accéder et d'observer un grand échantillon de données langagières. En utilisant les outils de la linguistique de corpus, et notamment les concordanciers, les apprenants recherchent certaines pratiques langagières ou certains mots puis observent et comprennent leur fonctionnement. L'accès à de grands corpus modifie l'accès aux compétences socio-interactionnelles par les apprenants. Ils observent, comparent, analysent le fonctionnement de la langue en cotexte et en contexte. Ils induisent eux-mêmes les règles, les usages et les variations sociolinguistiques des pratiques langagières. Lors de plusieurs expérimentations dans un département de FLE, ces démarches et processus ont été filmés et analysés notamment pour saisir les activités métacognitives mise en œuvre.

3 Résultats

Nous montrerons comment les apprenants réussissent, accompagnés ou non d'enseignants, à tirer profit des corpus pour apprendre à comprendre et à s'exprimer en interaction. Pour cela, nous présenterons les démarches d'observation, d'analyse et d'appropriation des apprenants travaillant principalement avec le corpus TCOF, mentionné précédemment. Nous illustrerons également ces démarches avec des expérimentations menées avec le corpus FLEURON, corpus

multimédia interrogeable par un concordancier multimodal et intégré à un dispositif numérique d'apprentissage du FLE (https://fleuron.atilf.fr/).

Nous examinerons plus particulièrement comment les apprenants réussissent à s'approprier des pratiques langagières telles que « du coup », « quand même » ou « genre » (en tant que marqueur interactionnel) en interaction. Nous verrons également comment les apprenants s'emparent de leurs observations des interactions pour comprendre les mécanismes de prise de paroles et les normes interactionnelles. Nous mettrons au jour leur démarche d'analyse inductive, tels des « Sherlock Holmes » ou des détectives de la langue (Johns 1997). Nous verrons que les compétences socio-interactionnelles acquises grâce à l'ASC permettent aux apprenants d'interagir de façon efficace, des points de vue linguistique, pragmatique, interactionnel et culturel (André 2017).

Références bibliographiques

Abe, D., Carton, F., Cembalo, S. M., et Régent, O. (1979). Didactique et authentique : du document à la pédagogie. *Mélanges pédagogiques*, 1-14. http://www.atilf.fr/spip.php?article3580

André, V. (2018). Nouvelles actions didactiques : faire de la sociolinguistique de corpus pour enseigner et apprendre à interagir en français langue étrangère. *Action didactique*, n°1, 71-88. http://univ-bejaia.dz/pdf/ad1/Andre.pdf

André, V. (2017). Un corpus multimédia pour apprendre à interagir en situations universitaires en France. *Actes du troisième colloque international de l' ATPF « Enseigner le français : s' engager et innover »*, 292-315. http://www.atpf-th.org/Actes_du_Colloque.pdf

André, V. (2015). Sociolinguistique des interactions verbales : de l' analyse des situations de travail aux implications sociales. *Langage, Travail et Formation*, 1-10. https://reseaultf.atilf.fr/wp-content/uploads/2015/10/Virgine-Andre.pdf

Aston, G. (Ed.). (2001). *Learning with corpora.* Athelstan.

Boulton, A., Cobb, T. (2017). Corpus Use in Language Learning : A Meta-Analysis. *Language learning.* Volume 67, Issue 2, 348-393. https://doi-org.bases-doc.univ-lorraine.fr/10.1111/lang.12224

Boulton, A., Tyne, H. (2014). *Des Documents Authentiques aux Corpus. Démarches pour l'Apprentissage des Langues.* Didier.

Conrad, S. (2000). Will Corpus Linguistics Revolutionize Grammar Teaching in the 21st Century ? *TESOL*, Vol 34, N°3, 548-560.

Duda, R., Esch, E., Laurens, J. P. (1972). Documents non didactiques et formation en langues. *Mélanges pédagogiques*, 1-48. http://www.atilf.fr/spip.php?article3528

Johns, T. (1991). Should you be persuaded : Two examples of data-driven learning. *In* T. Johns, P. King (dir.), *Classroom Concordancing. English Language Research Journal,* 4, 1-16.

Johns, T. (1997). Contexts : The background, development and trialling of a concordance-based CALL program. *In* Wichmann, A., Fligelstone, S., McEnery, T., Knowles, G. (dir.), *Teaching and Language Corpora.* Harlow : Addison Wesley Longman, 100-115.

Sinclair, J. (1996). Preliminary recommendations on corpus typology. *EAGLES Document TCWG-CTYP/P*. http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html

Timmis, I. (2015). *Corpus Linguistics for ELT. Research and Practice.* London : Routledge.

Session 2.B.
Analyses lexicologiques et diachronie

Le projet CONDÉ : présentation. Les défis d'un corpus de textes en diachronie longue

Mathieu Goux et Morgane Pica
CRISCO (EA4255) Université UNICAEN (Caen Normandie)

Le projet RIN CONDÉ (CONstitution d'un Droit europÉen : six siècles de coutumiers normands), financé par la région Normandie, se consacre à l'histoire de la Coutume de Normandie, jusqu'à présent surtout étudiée selon les circonstances de sa constitution, du XIII^{e} au XVIII^{e} siècles. Ce projet propose de considérer ces textes comme des témoignages précieux relatifs à l'histoire du droit, notamment dans les rapports à la jurisprudence et au pouvoir temporel ; du point de vue historique, le travail fait sur les versions et le nombre de manuscrits propose un panorama de la structuration de l'écrit et de son économie dans le domaine juridique ; enfin, ce matériau abondant est une source précieuse quant à une analyse linguistique diachronique, par l'intermédiaire de textes précisément datés et localisés.

CONDÉ édifie une base de données interrogeable en ligne, accessible à la communauté des chercheurs comme des professionnels du droit ou des enseignants de l'ancienne langue, reflétant l'évolution et l'élaboration du droit en Normandie. Ce projet demande à surmonter un certain nombre de difficultés : en amont, par l'inventaire et la sélection des coutumiers représentatifs ; puis par les défis qu'engagent la transcription, la numérisation et l'étiquetage syntaxique des textes ; enfin, par les diffusions potentielles d'un tel corpus, jamais envisagé sous cet angle jusqu'à présent.

Nous présenterons notre corpus sous ses aspects techniques, en nous inscrivant dans le champ contemporain des Humanités Numériques d'une part, de la linguistique de corpus de l'autre, et nous présenterons la façon dont nous avons retranscrit la spécificité de ce corpus, enrichi d'un appareillage scientifique impliquant (i) une numérisation et une transcription de ces différents témoins ; (ii) un système de renvoi et de navigation permettant un accès intelligent à leurs différents contenus ; (iii) une lemmatisation et un étiquetage morphosyntaxique, permettant de faire des recherches fines au sein des textes.

Ces différents aspects nécessitent de résoudre plusieurs défis d'ordre philologique et linguistique. Tout d'abord se pose la question de la matérialité des témoins, et la façon dont la base de données se doit à la fois de restituer la typo-disposition générale des documents tout en rendant la transcription accessible au plus grand nombre. Les coutumiers proposent effectivement une architecture interne spécifique, dont il convient de rendre compte : les articles de la coutume, généralement numérotés de façon suivie et organisés en grandes catégories, sont commentés et expliqués par les auteurs de la coutume. Cette glose est elle-même organisée par des systèmes complexes d'indications marginales, de notes de bas de page, de renvois divers à l'importance distincte dans la compréhension du propos. La transcription des témoins se doit ainsi de restituer cette hiérarchisation intrapaginale, en repensant l'architecture logique des manuscrits et des imprimés et en profitant au mieux des avantages offerts par le passage sur le support numérique.

Cette nouvelle façon de songer les témoins doit rendre non seulement compte de l'organisation interne de chaque document, mais également les penser dans leur continuité diachronique et leur histoire éditoriale. L'organisation de la base de données se doit effectivement de permettre une navigation scientifiquement pertinente dans l'ensemble des coutumiers, ce qui amène à repenser la progression du commentaire, mais également à envisager des continuités et des ruptures dans la façon dont les spécialistes du droit ont envisagé leur matériau. Partant, il ne nous suffira pas d'établir des éditions diplomatiques et/ou critiques de ces différents coutumiers, mais de les insérer dans un parcours représentatif de l'histoire moderne du livre, l'enrichissement scientifique proposé permettant de restituer au mieux les mouvements des pratiques éditoriales au long de l'histoire.

Afin d'assurer un accès facilité à cette abondante masse de données, et de naviguer au mieux entre les différents coutumiers, nous proposerons, en marge d'une organisation thématique, un outillage morpho-syntaxique par l'intermédiaire d'une lemmatisation et d'un étiquetage des parties du discours que nous autorise l'encodage XML-TEI. Cet enrichissement permettra d'opérer des recherches approfondies au sein des textes, soit par mots-clés, soit par catégorie grammaticale. Cette opération en appelant aux avancées du TAL, elle se confronte en particulier aux problématiques relatives à l'analyse automatique des discours en diachronie. Ces problématiques se ramifient dans plusieurs directions complémentaires : (i) d' une part, concernant la phase de lemmatisation, et la reconnaissance des formes présentes dans les coutumiers. Les logiciels de traitement automatique à la destination des chercheurs, à l'instar de LGeRM (ATILF) ou Lemming (DEAF) sont aujourd'hui robustes, et permettent une reconnaissance fine des occurrences qui se présentent à nous. Notre parcours des coutumiers permettra cependant d'enrichir les dictionnaires, notamment par l'étude de textes de spécialité dont le jargon présente des termes rares ou spécifiques que nous trouvons peu dans les textes littéraires ; (ii) d'autre part, par la reconnaissance des parties du discours. Une fois encore, les logiciels consacrés et développés depuis plusieurs années, au gré de plusieurs projets (voir bibliographie), ont atteint un taux de réussite satisfaisant que notre corpus permettra d'améliorer par l'étude de textes jusqu'à présent peu étudiés ; (iii) enfin, par le jeu d'étiquettes retenu pour conduire cette reconnaissance.

Ce dernier point sera sans doute le plus difficile à mettre en place, et celui qui nécessitera de notre part la plus grande réflexion. Nous devons effectivement sélectionner un jeu d'étiquettes conciliant plusieurs critères : tout d'abord, une opérabilité sur une période diachronique longue, puisque devant couvrir plus de six siècles d'évolution linguistique. Nous ne pouvons effectivement, à la fois pour des raisons autant pratiques que scientifiques, préparer un jeu d'étiquettes spécifique par période temporelle : cela rendrait non seulement l'encodage des documents particulièrement fastidieuse, mais compliquerait également indûment la recherche des occurrences et des structures dans les textes de notre corpus en multipliant les options de recherche. Nous ne pouvons cependant point reprendre tel quel le système retenu par les grammaires d'usage contemporaines, car la catégorisation *stricto sensu* opérée par celles-ci est inapte pour les états antérieurs de la langue ; du moins, elle demande d'élaborer des règles justifiant leur transfert et des choix scientifiques assurant l'interopérabilité des catégories modernes aux textes anciens. Ce transfert touche, par ailleurs, non seulement les parties du discours en elles-mêmes, mais

également l'identification de la catégorie du *mot*, dont les limites ont été au cours de l'histoire linguistique redéfinies au gré d'opérations de segmentation et de concaténation qu'il nous faudra expliciter au mieux.

Un autre écueil qu'il nous faudra affronter, est celui du grain de ce jeu d'étiquettes. S'il est tentant d'élaborer un ensemble de règles et d'éléments décrivant au mieux la spécificité de chaque occurrence du point de vue morpho-syntaxique et ce à chaque étape de l'évolution de la langue, un jeu d'étiquettes par trop complexe ralentirait nécessairement les recherches, et empêcherait les usagers de la base de données de naviguer efficacement parmi les occurrences. À l'inverse, un jeu d'étiquettes trop simplexe offrirait un grain trop grossier pour être scientifique viable, et ne permettrait point d'opérer des analyses fines des discours et d'en apprécier l'évolution diachronique. Un équilibre se doit donc d'être trouvé, équilibre devant prendre en compte, de plus, les moyens mis à notre disposition pour mener à bien ce projet, le temps que nous pouvons consacrer à cette phase de l'outillage, et les difficultés techniques présidant à sa mise en place. Une réflexion de fond, associant linguistique diachronique, humanités numériques, linguistique de corpus et histoire du livre, se cheville donc nécessairement à la conduite du projet CONDÉ, réflexion certes entée à la spécificité de notre objet d'étude mais dont les prolongations, les acquis et les solutions, permettront de mieux cerner la façon dont ces grands corpus diachroniques peuvent être restitués grâce aux outils numériques.

Notre présentation se focalisera ainsi sur les spécificités du corpus CONDÉ. Nous développerons notre propos dans trois directions complémentaires. (i) Dans un premier temps, nous présenterons les textes de notre « corpus noyau », soit les manuscrits et imprimés que nous avons retenus pour le projet CONDÉ et qui sont représentatifs non seulement de l'évolution de l'histoire de la coutume au long du temps, mais témoignent également des modifications des pratiques éditoriales et de l'histoire de la langue française. Nous justifierons les choix effectués des points de vue linguistiques et historiques, et les caractéristiques générales de leur typo-disposition éditoriale. (ii) Dans un second temps, nous présenterons les logiciels que nous avons exploités pour la numérisation et la transcription de nos témoins, les défis que nous avons dû relever ce faisant, et le jeu de balises XML-TEI que nous avons retenus pour l'encodage de la transcription. (iii) Enfin, le jeu d'étiquettes sur lequel nous nous sommes arrêtés pour l'analyse PoS (*Part of Speech*) du corpus, qui propose un étiquetage homogène de la langue des témoins sur une diachronie longue (près de six siècles), ainsi que les règles de segmentation que nous avons établies. Une fois encore, nous reviendrons sur les logiciels que nous avons exploités pour conduire cette phase d'enrichissement du corpus, les difficultés que nous avons rencontrées et la façon dont nous avons relevé les défis de ce projet d'une envergure unique.

Références bibliographiques

Aït Mokhtar, S., Chanod, J.-P. & Roux, C. (2001). Robustness beyond Shallowness : Incremental Deep Parsing. *Natural Language Engineering,* 8, 121-144.

Capin, D. & Larrivée, P. (2017). Gloses et réécritures des textes coutumiers : les métamorphoses de la Coutume de Normandie du Moyen Âge à la Renaissance. *Le français préclassique,* 19, 49-68.

Caron, P. (2002). Vers la notion de chronolecte. Quelques jalons à propos du français préclassique ». Sampson, R. & Ayres-Bennett, W. (éd.). *Interpreting the History of French. A Fetschrift for Peter Rickard on the occasion of his eightieth birthday.* Amsterdam : Rodopi. 329-352.

Démare-Lafont, S. & Lemaire, A. (dir.) (2010). *Trois millénaires de formulaires juridiques, Actes de la Table ronde des 28 et 29 septembre 2006, Hautes Études Orientales –Moyen et Proche-Orient,* 48. Genève : Droz.

Heiden S., Magué J.-P., Pincemin B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie. Conception et développement. Dans Bolasco S., Chiari, I., Giuliano, L. (eds). *JADT 2010 : 10th International Conference on Statistical Analysis of Textual Data.* Rome : Edizioni Universitarie di Lettere Economia Diritto.

Holmes, M. (2010). Using the Universal Similarity Metric to Map Correspondences between Witnesses (Conference presentation). *Digital Humanities 2010, Conference, Kings College London.*

Lay M.-H. et Pincemin B. (2010). Pour une exploration humaniste des textes : AnaLog. Dans Bolasco S., Chiari, I., Giuliano, L. (eds), *JADT 2010 : 10th International Conference on Statistical Analysis of Textual Data.* Rome : Edizioni Universitarie di Lettere Economia Diritto.

Neveux, F. (2011). Le contexte historique de la rédaction des coutumiers normands. *Annales de Normandie*, 61.2, 11-22.

Pigeon, J. (2015). Les Observations sur la Coutume de Normandie : vers une unification aux inspirations normandes. *Les Annales de droit,* 7, 177-201.

Pincemin B. (2004). Lexicométrie sur corpus étiquetés. Dans Purnelle, G. *et al.* (éds). *Actes des 7es Journées internationales d'analyse statistique des données textuelles (JADT 2004)*. Louvain-la-Neuve : Presse universitaires de Louvain. II, 865-873.

Pincemin B., Issac F., Chanove M., Mathieu-Colas M. (2006). Concordanciers : thème et variations. Dans Viprey J.-M. *et al.* (éds). *Actes des 8e Journées internationales d'Analyse statistique des Données Textuelles (JADT 2006).* Besançon : Presses Universitaires de Franche-Comté. II, 773-784.

Prévost S., Heiden S. (2002). Etiquetage d'un corpus hétérogène de français médiéval : enjeux et modalités. Dans Pusch C. D. & Raible W. (eds.). *Romanistische Korpuslingustik : Korpora und gesprochene Sprache / Romance Corpus Linguistics : Corpora and Spoken Language.* Tübingen : G. Narr. 127-136.

Étude métalexicographique diachronique de l'anglais : méthodes et outils pour des analyses lexicologiques et morphologiques

Sylvie Hanote [1], Michaël Nauge [1], Nicolas Trapateau [2] et Franck Zumstein [3]. [1]FoReLLIS, Université Poitiers
[2]Bases Corpus Langage UMR 7320, Université de Nice-Côte d'Azur
[3]CLILLAC-ARP, Université de Paris 7 - Denis Diderot
sylvie.hanote@univ-poitiers.fr, michael.nauge@univ-poitiers.fr, nicolas.trapateau@unice.fr, franck.zumstein@univ-paris-diderot.fr

1 Introduction

Cette première étude s'inscrit dans un projet plus large (Projet DicoDiachro[1]) qui explore la méthodologie d'analyse d'un corpus de dictionnaires orthoépiques consignant la prononciation de l'anglais britannique de 1700 à 1990. Une équipe multi-sites[2] de chercheurs propose une étude pluridisciplinaire à partir de sept dictionnaires[3] de l'anglais alliant les humanités numériques, la phonologie et la lexicologie historique. Les dictionnaires choisis ont l'intérêt de couvrir trois siècles mais ont aussi des tailles et des objectifs différents, ce qui induit des choix spécifiques faits par les lexicographes (présence ou non de définitions des mots, d'une représentation de leur prononciation et d'éventuelles variantes, des remarques de la part des auteurs, etc.) qu'il est intéressant de mettre au jour. Dans le cadre de cette étude spécifique, nous nous proposons de présenter la méthodologie générale du projet en cours et les résultats liés à la couverture lexicale (les apparus et les disparus) entre trois dictionnaires : Buchanan (1766) et Walker (1791) pour la deuxième moitié du 18^{me} siècle, et le dictionnaire de J.C. Wells (1990) pour le « point de vue » contemporain et nous nous concentrons sur un point d'entrée spécifique, les préfixés, pour ce qui concerne l'analyse qualitative des données.

2 Corpus et méthodologie

2.1 Corpus

Dans le cadre du projet DicoDiachro piloté à l'Université de Poitiers, tous les dictionnaires que nous avons retenus permettent de couvrir la période qui va de l'anglais moderne tardif jusqu'à l'anglais actuel, depuis les premiers dictionnaires marquant systématiquement la place de l'accent (Bailey 1727) en passant par ceux qui introduisent des transcriptions des formes orales des mots-vedettes (Buchanan 1766, Jones ca. 1805). Le dictionnaire de Walker (1791) ajoute à cela des remarques sociolinguistiques et des règles de prononciation, et le dictionnaire encyclopédique de Wright (1852-1856) permet d'avoir une couverture lexicale et phonétique la plus large possible de la période du 19^{me} siècle. Parmi ces dictionnaires, seuls ceux de Buchanan et Walker

1. Le projet DicoDiachro a été lauréat d' un appel à projet 'Humanités numériques' de la MSHS de Poitiers en 2017.

2. Les universités partenaires du projet sont les suivantes : Côte d'Azur, Poitiers, Paris, Picardie - Jules Verne.

3. ailey (1727), Barclay (1774), James Buchanan (1766), Stephen Jones (ca. 1805), John Walker (1791 [1809]), Wright (1852-1856), J.C. Wells (1990).

ont bénéficié d'une numérisation complète à ce jour et ont déjà permis de mener recherches doctorales sur l'évolution de la prononciation du lexique anglais (Castanier, 2016 ; Trapateau, 2015). C'est à partir leurs données que nous tirons les conclusions métalexicographiques de cette étude. Les deux dictionnaires choisis pour l'étude sont presque contemporains et s'ils présentent une différence de volume (le dictionnaire de Buchanan est bâti sur une nomenclature qui compte 26 229 entrées alors que le dictionnaire orthoépique de Walker comporte 38 774 entrées), il nous est apparu que les deux bases de données qu'ils offrent sont commensurables et justifient un travail de comparaison susceptible de manifester, à une génération d'écart environ, une évolution des nomenclatures qui reflète des changements orthographiques, phonétiques, accentuels et morphologiques et de nous donner des indications sur les choix des lexicographes. Les données issues de ces deux dictionnaires sont enfin mises en regard de celles du dictionnaire de J.C. Wells (1990) pour le « point de vue de la langue contemporaine ».

2.2 Méthodologie de création d'un corpus de dictionnaires comparables

Afin de mener des analyses exhaustives et outillées de nos dictionnaires, nous avons mis en œuvre plusieurs étages de transformation des données. Notre objectif est de disposer de nos dictionnaires sous un format pivot permettant des interrogations croisées malgré la diversité des dictionnaires choisis.

Pour chaque dictionnaire, après le parcours classique de numérisation, OCR, correction d'OCR (et balisage éventuel), nous introduisons une étape de *parsing* pour restructurer les données sous forme matricielle : chaque ligne contient une entrée et chaque colonne un type de variable descriptive (Table1).

HWD	POS	Orig.	Def.
A′BBACY	S.	L.	the rights and privileges of an abbot.
A′BBESS	S.		a governeſs of nuns.
TO A′BDICATE	V.ACT.	L.	to give up a right.

Tab. 1 : Exemple de matrice générée –Dict. Walker

Il est alors possible de lancer des analyses avec des outils courants comme MS-Excel, Libre Office Calc ou des scripts R, Python ou autre, pour des recherches reproductibles.

Nous introduisons également un étage de contrôle qualité sur chaque colonne pour identifier, entre autres, les occurrences de valeurs vides ou de variations inattendues. Nous avons ainsi identifié des erreurs d'OCR non corrigées, des erreurs d'impressions ou des variantes d'auteur.

Dans le cas d'analyses croisées avec plusieurs dictionnaires nous avons ajouté une étape de normalisation pour rendre les entrées lexicales et les champs descriptifs comparables au sens informatique du terme et produire une nouvelle édition électronique dite normalisée. Pour cela, nous définissons des règles formelles pour l'entrée et ses champs descriptifs. Pour les entrées,

nous avons spécifié une casse, supprimé *to* devant les verbes dans les dictionnaires qui les codaient ainsi, etc. Pour le champs POS, nous définissons un appariement/alignement entre les valeurs du dictionnaire traité et les valeurs d'un vocabulaire de référence comme peut l'être un thésaurus ou termweb s'appuyant lui même sur les références ISO 12620 et ISO 30042. L'important ici n'est pas le vocabulaire de référence choisi, mais plutôt l'idée d'un pivot commun permettant à tous les dictionnaires d'utiliser un même vocabulaire et de rendre ainsi comparable les V. du dictionnaire A au Verbe du dictionnaire B :

- Les valeurs POS *V.* du dictionnaire_A deviennent *Verb* dans le dictionnaire_A_normalisé
- Les valeurs POS *Verbe* du dictionnaire_B deviennent *Verb* dans le dictionnaire_B_normalisé

Ainsi, bien qu'initialement très différents, nos dictionnaires une fois normalisées peuvent répondre de manière homogène à nos requêtes. Nous pouvons alors initier nos analyses croisées diachroniques.

2.3 La normalisation des données : une première question de variation (orthographique)

La détection des mots apparus / disparus entre les dictionnaires nous a conduit à prendre en compte la variation orthographique entre les dictionnaires de référence. Avec des *a priori* et des connaissances scientifiques, il est aisé de prévoir une partie des variantes orthographiques (par exemple la variation entre *-ic* et *-ick* à la finale des noms et adjectifs empruntés au français ou au latin), mais que faire de celles qu'on ne peut pas anticiper ?

2.4 Premiers résultats et discussion

Ce premier travail sur la normalisation des données nous a permis d'obtenir la densité lexicale pour chacun des dictionnaires du corpus : nombre d'entrées dans le dictionnaire (cf. axe des ordonnées) par lettre (en abscisse), comme indiqué dans la figure 1 ci-dessous.

Une recherche complémentaire s'impose prenant en compte l'inégale densité du lexique pour chaque lettre du dictionnaire, problématique d'autant plus cruciale que les techniques d'imprimerie d'alors conduisaient les lexicographes à rédiger leurs dictionnaires de la lettre A à la lettre Z, selon une approche chronologique, ce qui n'est pas le cas du dictionnaire contemporain de Wells.

2.5 Vers une analyse qualitative des données : un début d'étude métalexicographique

À partir d'une étude comparative systématique des dictionnaires de Buchanan et Walker, il est apparu que le plus grand nombre d'entrées lexicales dans le dictionnaire de Walker est lié à des choix de l'auteur qu'il nous est apparu intéressant d'étudier : inclusion des unités lexicales construites à partir des génitifs (génitifs de mesure) et des formes composées, recours à un fonds gréco-latin important, prise en compte de la dérivation (préfixation, suffixation), par exemple. Nous nous intéressons plus précisément à la préfixation. Dans la figure 2 ci-dessous, on remarque

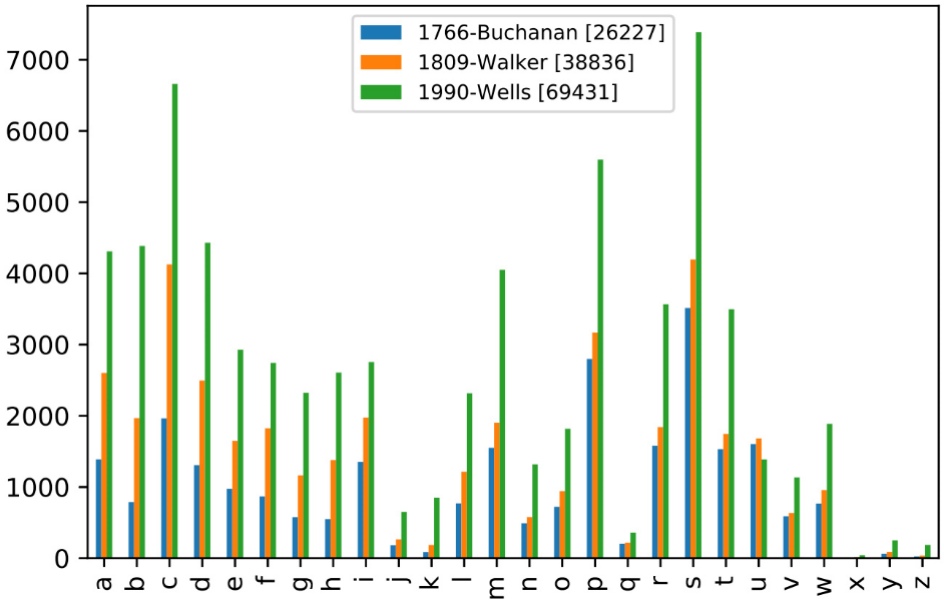


FIG. 1 : Nombre d'entrées lexicales par lettre sur les trois dictionnaires du corpus d'étude

que les mots préfixés en *dis-* sont deux fois plus nombreux dans le dictionnaire de Walker que dans celui de Buchanan, et la présence croissante des préfixés en *de-*. À l'inverse, il faut noter le déclin des préfixés en *un-* entre la seconde moitié du 18^{me} siècle et l'époque contemporaine. Il nous faudra aborder les questions de productivité (Tournier, 1985 ; Trevian, 2010 ; Säily, 2018) et de rivalités entre préfixes, à l'instar des suffixes (Arndt-Lappe, 2014).

Ces résultats seront présentés et discutés aussi bien en synchronie large (étude comparée des données obtenues à partir des dictionnaires de Buchanan et Walker en prenant en compte le phénomène de dérivation mais aussi l'influence des langues romanes, notamment du français) qu'à partir du point de vue contemporain.

En ce qui concerne la perspective synchronique, l'étude présentée ne peut pas faire l'économie d'une contextualisation socio-historico-linguistique des dictionnaires à l'étude (Säily, 2014) et d'une discussion autour des observations et des choix des lexicographes anglais de l'époque, différents de ceux des français dont la longue tradition lexicographique a permis les neufs éditions du dictionnaire de l'Académie française du 17^{me} siècle à nos jours (Quémada, 1997).

En ce qui concerne le « point de vue contemporain », nous démontrerons l'intérêt d'un traitement complémentaire des données en intégrant l'utilisation de Tree-Tagger[4] pour faire état des entrées aujourd'hui considérées comme obsolètes ou du moins méconnues. Enfin, pour compléter le panorama de l'évolution du lexique sur la période qui va de l'anglais moderne tardif jusqu'à l'anglais actuel, un long travail de numérisation et normalisation des données reste bien entendu

4. Modèle english-bnc.par.gz construit à partir du British National Corpus.

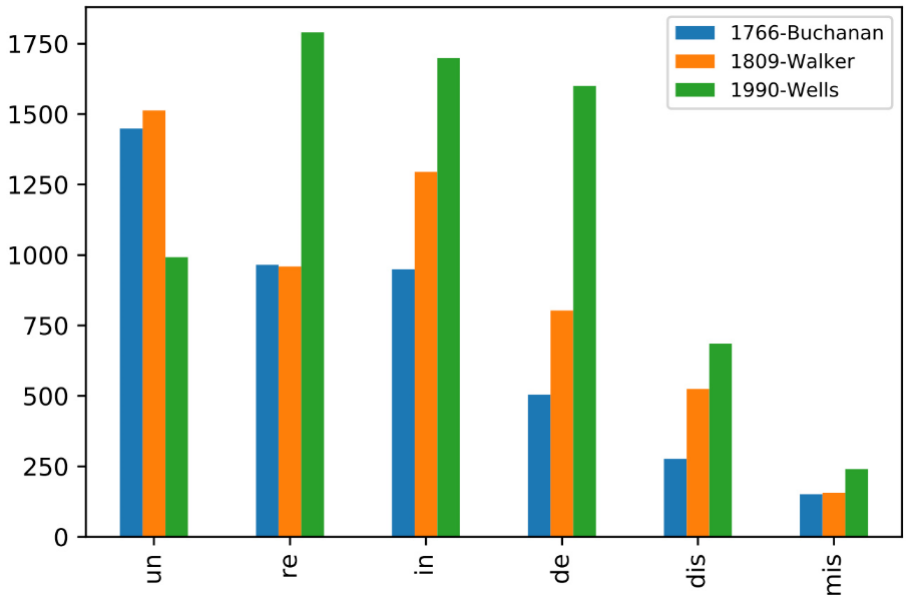


FIG. 2 : Nombre d'entrées lexicales par préfixe sur les trois dictionnaires du corpus d'étude

à poursuivre dans le cadre du projet global (Trapateau *et al*, 2019).

Références bibliographiques

Arndt-Lappe S. (2014). Analogy in suffix rivalry : the case of English -ity and -ness, *English Language and Linguistics* 18.3 : 497–548

Castanier, J. (2016). *L'évolution accentuelle du lexique en anglais contemporain appréhendée à travers les dictionnaires de prononciation (18^{me} –21^{me} siècles)*. Thèse de doctorat non publiée, Université de Poitiers.

Quémada, B. (ed.). (1997). *Les préfaces du dictionnaire de l' Académie française 1694-1992,* Paris, Honoré Champion.

Säily, T. (ed.). (2014). *Sociolinguistic variation in English derivational productivity : Studies and methods in diachronic corpus linguistics,* Helsinki, Société Néophilologique.

Säily, Tanja, (2018), Change or variation ? Productivity of the suffix *-ness* and *-ity*, Terttu Nevalainen, Minna Palander-Collin & Taina Säily (eds.), *Patterns of Change in 18th-century English : A Sociolinguistic Approach*. Amsterdam : John Benjamins, p. 197-218.

Tournier, J. (1985). *Introduction descriptive à la lexicogénétique de l'anglais contemporain*, Paris –Genève, Champion - Slatkine.

Trapateau, N. (2015). *Placement de l'accent et voyelles inaccentuées dans la prononciation de l'anglais du XVIIIe siècle sur la base du témoignage des dictionnaires de prononciation, des vers et de la musique vocale*, Thèse de doctorat non publiée, Université de Poitiers.

Trapateau, N. Videau, N. Duchet, J.L. et Hanote, S. (2019). Quelles méthodologies pour constituer et exploiter des corpus de données orales anciennes et contemporaines ?, in Caron, P. Defiolle, R. et Lay, M.H. (eds.). *Données, métadonnées des corpus et catalogage des objets en sciences humaines et sociales*, Rennes, PUR, 171-188.

Trevian, I. (2010). *Les affixes anglais, productivité, formation de néologismes et contraintes combinatoires : De la diachronie à la synchronie*, Bern, Peter Lang.

Sources primaires :

Bailey, Nathan, *An Orthographical Dictionary, shewing both the orthography and the orthoepia of the English tongue*, London : T. Cox, 1727.

Barclay, James, *A Complete and Universal English Dictionary on a New Plan*, London : Richardson et al., 1774.

Buchanan, James, *An Essay Towards Establishing a Standard for an Elegant and Uniform Pronunciation of the English Language, throughout the British dominions, as practised by the most learned and polite speakers,* London : E.N.C. Dilly, 1766.

Jones, Stephen, *Sheridan improved, A general pronouncing and explanatory dictionary of the English language*, The Tenth edition, London : Vernor and Hood, 1805.

Walker, John, *A critical pronouncing dictionary, and expositor of the English language*, 1791, 6th stereotype edition, 1809, 26th stereotype edition, 1823 ; 29th stereotype edition, 1827.

Wells, John Christopher, *Longman Pronunciation Dictionary*, 1st edition, 1990.

Wright, Thomas, *The Universal Pronouncing Dictionary and General Expositor of the English Language*. London : J. & F. Tallis, 1852-1856.

Les terminaisons -ic et -ical en anglais : essai de comparaison métalexico-graphique entre les dictionnaires de Buchanan (1766) et de Walker (1791)

Jean-Louis Duchet [1], Franck Zumstein [2], Nicolas Trapateau [3], Sylvie Hanote [1] et Michael Nauge [1].

[1]FORELLIS, Université de Poitiers

[2]CLILLAC, Université Paris 7- Denis Diderot

[3]UMR 7320 BCL, Université Côte d' Azur

jlduchet@univ-poitrs.fr, franck.zumstein@univ-paris-diderot.fr, sylvie.hanote@univ-poitiers.fr, michael.nauge@univ-poitiers.fr

1 Introduction

Cette contribution se propose d'illustrer, par une étude de cas, un projet multi-site de comparaison de dictionnaires de prononciation. L'*Essay* de Buchanan de 1766 est bâti sur une nomenclature qui compte 26 229 entrées. Le dictionnaire orthoépique de Walker, le *Critical Pronouncing Dictionary and Expositor of the English Language* de 1791 comporte 38 774 entrées. En dépit de cette différence de volume, il nous est apparu que les deux bases de données phonétiques qu'ils offrent sont commensurables et justifient un travail de comparaison susceptible de manifester, à une génération d'écart environ, une évolution de la nomenclature susceptible de refléter des changements orthographiques, phonétiques, accentuels, mais aussi morphologiques, sur lesquels nous souhaiterions attirer l'attention des linguistes. Au-delà des travaux antérieurs sur les problématiques de compétition suffixale (Nurmi 1996, Säily, 2008 & 2014, et Arndt-Lappe 2014), nous voudrions montrer l'intérêt d'une comparaison inter-dictionnaires construite par soustraction des entrées à partir de versions numérisées des deux dictionnaires afin de soulever les problèmes de normalisation orthographique et lexicographique et de différence de couverture lexicale.

2 Corpus et méthodologie

2.1 Corpus

Afin de mener des analyses exhaustives et outillées de nos dictionnaires, nous avons mis en œuvre plusieurs étages de transformation des données. Notre objectif est de disposer de nos dictionnaires sous un format pivot permettant des interrogations croisées malgré la diversité des dictionnaires choisis.

Pour chaque dictionnaire, après le parcours classique de numérisation, OCR, correction d'OCR (et balisage éventuel), nous introduisons une étape de parsing pour restructurer les données sous forme matricielle : chaque ligne contient une entrée et chaque colonne un type de variable descriptive.

2.2 Méthodologie

Il est alors possible de lancer des analyses avec des outils courants comme MS-Excel, Libre Office Calc ou des scripts R, Python ou autres, pour des recherches reproductibles. Nous introduisons également un étage de contrôle qualité sur chaque colonne pour identifier, entre autres, les occurrences de valeurs vides ou de variations inattendues. Nous avons ainsi identifié des erreurs d'OCR non corrigées, des erreurs d'impressions ou des variantes d'auteur.

Dans le cas d'analyses croisées avec plusieurs dictionnaires nous avons ajouté une étape de normalisation pour rendre les entrées lexicales et les champs descriptifs comparables au sens informatique du terme et produire une nouvelle édition électronique dite normalisée. Pour cela, nous définissons des règles formelles pour l'entrée et ses champs descriptifs. Pour les vedettes d'entrées, nous avons spécifié la casse, supprimé *to* devant les verbes dans les dictionnaires qui les codaient ainsi, etc. Pour le champs POS, nous définissons un alignement entre les valeurs du dictionnaire traité et les valeurs d'un vocabulaire de référence comme peut l'être un thésaurus ou termweb s'appuyant lui même sur les références ISO 12620 et ISO 30042. L'important ici n'est pas le vocabulaire de référence choisi, mais plutôt l'idée d'un pivot commun permettant à tous les dictionnaires d'utiliser un même vocabulaire et de rendre ainsi comparable les V. du dictionnaire A au Verbe du dictionnaire B :

- Les valeurs POS V. du dictionnaire_A deviennent Verb dans le dictionnaire_ A_normalisé
- Les valeurs POS Verbe du dictionnaire_B deviennent Verb dans le dictionnaire_B_normalisé

Ainsi, bien qu'initialement très différents, nos dictionnaires une fois normalisées peuvent répondre de manière homogène à nos requêtes. Nous pouvons alors initier nos analyses croisées diachroniques.

3 Résultats

Les mots terminés par *-ic* chez Buchanan

Le nombre d'occurrences se monte à 308. Si on examine ces mots en *-ic* qui trouvent une correspondance orthographique stricte dans Walker, on trouve un nombre d'occurrences limité à 5. En voici la liste :

N° ligne	Entrée	Buchanan 1766			Walker 1791		
		Part of speech	Transcription	Stress pattern	Part of speech	Transcription	Stress pattern
16	árabic	adjective	ărăbĭk	1-0-0	adjective	[a4r′-a4-bi2k]	1-0-0
30	balsámic	adjective	balsămĭk	0-1-0	adjective	[ba4l-sa4m′-i2k]	0-1-0
106	hieroglýphic	noun/ adjective	heerŏglĭfĭk	0-0-1-0	noun	[hi1-e1-ro1-gli2f′-fi2k]	0-0-0-1-0
254	specífic	adjective	speĕseĕfĭk	0-1-0	noun	[spe1-si2f′-i2k]	0-1-0
286	tónic	adjective	tŏnĭk	1-0	adjective	[to4n′-i2k]	1-0

TAB. 1 : Mots en *-ic* apparaissant sous la même forme chez Buchanan et Walker

Le schéma accentuel du dictionnaire de Buchanan est indiqué dans l'orthographe et non dans la transcription phonétique comme le fait Walker. Nous avons donc introduit dans la colonne « Stress pattern » une représentation normalisée de l'accent de mot pour faciliter la comparaison.

Les mots terminés par *-ic* chez Walker

Si on recherche les mots en *-ic* dans le dictionnaire de Buchanan qui ne trouvent *pas* de correspondance dans Walker, le nombre d'occurrences se monte à 303. La première raison en est la variation orthographique entre *-ic* et *-ick* à la finale des adjectifs empruntés au français ou au latin. Cette orthographe prônée par Johnson (1755), et reprise par Walker, vise à orthographier tous les adjectifs anglais terminés en *-ic* comme les mots monosyllabiques du fonds germanique (*thick, stick, etc.*) On comprend alors pourquoi *specific* est décrit comme nom par Walker : celui-ci distingue le nom *specific*, terme médical et pharmaceutique restreint, et l'adjectif *specifick*, à valeur sémantique large.

Si on ajoute la lettre <*k*> à ces 303 mots en *-ic* de Buchanan pour tenir compte de l'orthographe de Walker, on peut s'attendre à retrouver une correspondance beaucoup plus grande : de fait on en trouve plus de 300. Mais on constate néanmoins qu'il y a des mots de la nomenclature de Buchanan qui, même avec une finale en *-ick*, ne trouvent pas de correspondance dans Walker. Le nombre d'occurrences se monte à 29. En voici la liste :
anacreontic, antalgic, antartic, antiscorbutic, cholic, chylopoetic, diagraphic, dornic, drastic, homocentric, hypogastic, nosopoetic, odontalgic, pentastic, phagedenic, phonocamptic, phrentic, physiognomic, platic, poristic, postic, protatic, public-spirited, saic, sciatheric, selenographic, siccific, smegmatic, spagyric.

Certains de ces mots peuvent avoir disparu de l'usage, mais ce n'est certainement pas le cas de tous. Si on analyse les raisons possibles de la non-correspondance entre les deux nomenclatures de Buchanan et de Walker, on constate des faits disparates mais tous instructifs linguistiquement et culturellement et dont nous citons ici quelques extraits.

- *antiscorbutic* n'existe pas dans Walker mais l'adjectif *ANTISCORBUTICAL* y est attesté, ce qui va nous amener à examiner dans quelle mesure les adjectifs en *-ical* se substituent aux adjectifs en *-ic*, ou coexistent avec eux, s'en différencient sémantiquement, etc.

- *antartic* existe quand même dans Walker mais avec un *c* devant le *t* préfinal : *ANTARCTICK*, ce qui va nous amener à examiner les changements orthographiques autres que la variation entre *-ic* et *-ick*. De même, si *protatic*, dérivé de *protasis*, apparaît légitimement chez Buchanan, il n'en reste pas moins que Walker atteste de l'existence de *protactick*, clairement dérivé (par le sens décrit) de *protasis*, lui aussi.

- *hypogastic* est le produit d'une erreur d'orthographe de Buchanan. Cet adjectif se trouve chez Walker avec l'orthographe *hypogastrick* avec un *r* en préfinale, alors que par ailleurs Buchanan n'a pas *hypogastric*. On peut supposer que *antartic*, examiné ci-dessus, est le

fait d'une erreur semblable ou bien encore le reflet d'une prononciation dans l'usage réel qui procède à l'effacement du son [k] dans un agrégat de deux occlusives après [r] peu aisé à assimiler en anglais.

- L'entrée *dornic* n'est pas celle d'un adjectif en *-ic*, mais c'est un nom qui désigne le « drap de Tournai », qui se trouve sous l'orthographe *darnix* dans quelques rares dictionnaires de l'époque (le dictionnaire bilingue de Boyer & al., par exemple), et dans le dictionnaire de Wright (1852) sous l'orthographe *dornoch*, par réanalyse du toponyme écossais Dornoch, ville où est aussi produite une toile de nappe).

- *cholic* existe dans Walker, mais sans *h* après le *c* initial : *colick*. Il y a également dans le dictionnaire de Walker une entrée avec le *h*, mais qui n'est en l'espèce qu'une simple référence renvoyant à l'entrée *colic* : « Cholick.—See Colic ». Non suivie de la virgule habituelle, cette vedette avait échappé à un premier relevé. La double ponctuation avec le point et le tiret cadratin est utilisée régulièrement dans ce type de renvoi, et l'édition numérique du dictionnaire que nous utilisons respecte cette régularité et cette conformité à l'original. Ceci nous conduit à adopter des versions différentes des fichiers électroniques des dictionnaires orthoépiques, l'une strictement analogique, les autres diversement normalisées en documentant les critères de normalisation adoptés.

- *phrenetic* et *phrentic* apparaissent comme deux variantes chez Buchanan. Mais si *phrentic* n'apparaît pas sous cette forme chez Walker c'est parce que ce dernier l'orthographie *frantick*, différenciant ce terme à l'usage large du terme scientifique *phrenetic*, avec un *ph-* initial qui le caractérise nettement comme savant : l'orthographe et le registre sont solidaires.

- Les néologismes de l'époque sont diversement accueillis par les lexicographes et ceci explique qu'*antalgic* ne soit pas répertorié par Walker, alors qu'il l'était par l'encyclopédie de Chambers dès 1753 et le dictionnaire d'Ash en 1775. Buchanan semble réceptif à la néologie.
  Buchanan accepte *chylopoetic*, auquel Walker oppose *chylifactive* et *chylificatory*, de même sens, et que Buchanan n'inclut pas bien qu'il ait *chylifaction*. L'usage actuel, attesté par l'Oxford English Dictionary, donne raison à Buchanan et qualifie de 'rares' les néologismes reçus par Walker. Buchanan reçoit *drastic*, alors spécialisé dans le sens médical (« purgatif ») que Bentham et Stuart Mill au XIXe s. étendront à l'économie.
  Buchanan se montre encore plus réceptif que Walker au vocabulaire de l'astronomie avec *sciatheric*. Réceptif à la néologie, Buchanan donne les entrées apparentées *siccate, siccation, siccific, siccity,* là où Walker ne retient que le dernier de ces quatre termes relatifs à la sécheresse, au demeurant tous disparus de nos jours ; et s'il a les dérivés *desiccative* et *desiccation* il ne propose pas le verbe dérivant *desiccate* qui apparaît chez Buchanan. De même Walker accueille *physiognomer*, *physiognomy*, *physiognomist*, mais omet *physiognomic* alors que Buchanan retient les quatre et se montre plus systématique dans la prise en compte de la dérivation pour le choix de la nomenclature. Et si Walker retient bien *selenography*, il n'en tire pas l'adjectif *selenographic*, ce que Buchanan a en revanche prévu.

- La sensibilité du lexicographe au tabou peut jouer un rôle : *anacreontic* s'appliquant à des poèmes érotiques est présent chez Buchanan et absent chez Walker. Nous découvrons donc des différences métalexicographiques entre les dictionnaires du fait de leurs choix de nomenclature, quelques différences orthographiques et dérivationnelles, mais tout de même une convergence importante.
  Si, en effet, nous considérons maintenant les mots en *-ic* qui trouvent bel et bien une correspondance en *-ick* dans Walker, on obtient un nombre d'occurrences qui se monte à 274, parmi lesquels on trouve des entrées attendues : *civick, classick, comick, conick, critick, cynick, cholerick, despotick, eclectick, characteristick*. Mais on trouve aussi des adjectifs qui ne figurent plus dans les dictionnaires actuels à la nomenclature comparable à celle de Walker (ex. : *diarrhoetick*, qui est présent dans l'OED, mais sans attestation postérieure à 1880). Un adjectif comme *climacterick*, avec une orthographe sans *-k* final, existe encore : c'est l'un des rares adjectifs en *-ic* qui ne contraint pas l'accent principal sur la syllabe pénultième mais sur l'antépénultième : klaɪˈmæktərɪk (une variante américaine régularisée ˌklaɪmækˈterɪk est citée par Wells mais démentie par son CD- ROM dans l'exemple prononcé par un anglophone américain). Quant à *choleric*, transcrit [kólĕrĭk] par Buchanan en 1766, il est toujours accentué sur la même syllabe par Wells en 1990 : [ˈkɒlərɪk], mais avec une seconde variante régularisée avec un accent pénultième.
  Comme on peut s'y attendre vu leur date de publication les dictionnaires n'ont pas *electrick*, mais cette raison référentielle n'est pas, et de loin, la raison principale des différences constatées dans les nomenclatures comparées.

Si on examine maintenant les adjectifs en *-ical* dans le dictionnaire de Buchanan, on constate que leur nombre s'élève à 265.
Sur cet effectif, 241 (plus de 90%) se trouvent aussi dans la nomenclature de Walker. Il y en a toutefois 24 qui n'ont pas de correspondant chez Walker. En voici la liste : acronical, helispherical, horological, intrinsical, metropolitical, pancratical, paradisaical, pedagogical, planimetrical, prolifical, quadrinomical, quodlibetical, sciatherical, selenographical, simonical, sodomitical, sporadical, stalactical.
Dans cette liste, certains adjectifs se retrouvent chez Walker mais avec la terminaison *-ick*. Ainsi, au *prolifical* de Buchanan correspond le *prolifick* de Walker et son adverbe *prolifically*.
Il y a donc une compétition entre les deux terminaisons adjectivales dont témoignent les nomenclatures des dictionnaires. Nous couvrons là le terrain des conflits dérivationnels qui font l'histoire de la morphologie et étudiés par Säily et Arndt-Lappe.
L'adjectif *intrinsical* de Buchanan se retrouve dans la nomenclature de Walker en tant que *intrinsecal* (avec l'adverbe *intrinsecally*), et cette orthographe est fidèle à l'étymologie (le français *intrinsèque* n'est pas un suffixé en *-ique* et le latin a aussi un *e*). Mais le lexicographe de référence du XVIIIe siècle anglais, Samuel Johnson admet que l'orthographe *intrinsical*, que nous pouvons interpréter comme reconstruite à partir de la prononciation réduite à [ɪ] de la voyelle *e* inaccentuée, est devenue plus courante. Buchanan se rallie à cet usage, tandis que Walker, ici plus soucieux de fidélité à la langue savante, maintient une orthographe étymologique, d'autant plus facilement que la prononciation est la même pour les deux orthographes.
Enfin il y aussi des choix lexicographiques en partie arbitraires : Walker n'a pas *horological*,

mais il a *horology*, *horography* et *horometry*, tandis que Buchanan a *horological, horology* et *horometry*.
L'étude métalexicographique exhaustive des entrées exploitées comme base de données permet de dater les apparitions lexicographiques de mots nouveaux et de termes savants, de dépister la variation orthographique et phonétique (grâce à la rubrique orthoépique très présente dans les dictionnaires des XVIIIe et XIXe es. parce que recherchée par les lecteurs), de cerner les mouvements diachroniques de la morphologie dérivationnelle et même de dessiner la personnalité des lexicographes agissant comme prescripteurs autant que descripteurs de leur état de langue.

Références bibliographiques : sources primaires

Ash, John (1775). *The New and Complete Dictionary of the English Language.* London : for Edward and Charles Dilly and R. Baldwin, 1775.

Boyer, A., L. Chambaud, J. Garner (1829). *Dictionnaire anglais-français et français-anglais,* t. 1 Anglais- français. Paris, Ledentu, 1829.

Buchanan, James (1766). *An Essay Towards Establishing a Standard for an Elegant and Uniform Pronunciation of the English Language, throughout the British dominions, as practised by the most learned and polite speakers.* London : E.N.C. Dilly, 1766.

Johnson, Samuel (1755). *A General Dictionary of the English Language,* `https://johnsonsdictionaryonline.com/`)

Walker, John (1791, 6th stereotype edition, 1809). *A critical pronouncing dictionary, and expositor of the English langage*. London.

Wells, John C. (1990, 3rd ed 2008). *Longman Pronunciation Dictionary.* Harlow : Longman.

Wright, Thomas (1852-56). *The Universal Pronouncing Dictionary, and General Expositor of the English Language.* London, Edinburgh and Dublin : J & F Tallis, 5 vol.

Références bibliographiques : sources secondaires

Arndt-Lappe, Sabine (2014). Analogy in suffix rivalry : the case of English *-ity* and *-ness*. *English Language and Linguistics* 18.3 : 497–548.

Nurmi, Arja (1996). Periphrastic do and be + ing : Interconnected developments ? In Terttu Nevalainen & Helena Raumolin-Brunberg (eds.), *Sociolinguistics and language history : Studies based on the Corpus of Early English Correspondence* (Language and Computers : Studies in Practical Linguistics 15), 151–165. Amsterdam : Rodopi.

Säily, Tanja (2008). *Productivity of the Suffixes* -ness *and* -ity *in 17th-century English Letters : A Sociolinguistic Approach.* Thèse de Master, Department of English, University of Helsinki.

Säily, Tanja (2014). *Sociolinguistic variation in English derivational productivity : Studies and methods in diachronic corpus linguistics.* Mémoires de la Société Néophilologique de Helsinki, 284 p.

Session 3.A.
Cohérence, co-référence

De la coréférence exacte à la coréférence complexe : une typologie et sa mise en œuvre en corpus

Marine Delaborde et Frédéric Landragin
Lattice, CNRS, ENS, Université de Paris 3, PSL Research University, USPC
marine.delaborde@ens.fr, frederic.landragin@ens.fr

L ' annotation de la coréférence est une tâche délicate dès lors que l'on rencontre des expressions qui semblent référer à la même entité sans être exactement coréférentes. C'est un problème qui a été soulevé notamment par Recasens et al. (2011) pour le concept de near identity avec une catégorisation de différents types de relations de coréférence dans lesquelles les référents sont proches sans être exactement les mêmes.

Lorsque plusieurs expressions référentielles désignent le même référent, elles sont coréférentes et forment une chaîne de coréférence. Il s'agit principalement de noms communs, de noms propres et de pronoms. Les chaînes de coréférence peuvent s'étendre du début à la fin d'un texte, comme pour un personnage principal dans un roman par exemple, mais elles sont le plus souvent courtes et ponctuelles. Pour qu'il y ait référence, il faut pouvoir identifier un référent qui existe dans le monde ou que l'on peut se représenter (Charolles, 2002). Cependant, il arrive parfois que le référent d'une expression soit difficile à identifier de manière précise. Dans ce cas, la coréférence avec une autre expression référentielle dont on a identifié clairement le référent ne peut pas être stricte. C'est le cas par exemple des anaphores à antécédent flou, à ne pas confondre avec de l'ambiguïté (Fuchs, 1996).

> « Elle parle aussi avec une sentimentalité criante. Ma sœur et moi on l'arrête. On l'arrête à temps. Alors elle dit on ne me laisse pas parler ici. Mais ce ne sont pas des paroles qu'on a envie d'entendre, je ne sais pas pourquoi. »
> AKERMAN Chantal, Ma mère rit, 2013

Dans cet exemple, les deux premiers « on » coréfèrent et renvoient à l'antécédent « Ma sœur et moi » de manière évidente et stricte. En revanche, le troisième « on », qui est du discours rapporté, et le dernier « on » coréfèrent chacun de manière floue aux deux premiers car ils peuvent très bien renvoyer au même antécédent. Cependant le doute persiste en raison du fait qu'ils peuvent tous les deux avoir aussi une valeur générique. Il n'est donc pas raisonnable d'en faire une seule chaîne, mais les dissocier totalement ferait perdre une information.

L'annotation de la coréférence en corpus implique de faire des choix, notamment au niveau des relations. Le projet ACE (Doddington et al., 2004) distingue 5 types de relations entre les mentions : rôle, partie, localisation, proche et sociale. Le projet OntoNotes (Pradhan et al., 2011) différencie la coréférence identique de la coréférence appositive. C'est aussi le cas dans le corpus WikiCoref (Ghaddar & Langlais, 2016) qui distingue les coréférences identiques, attributives et attributives dans des constructions copulatives. Dans le corpus Phrase Detectives (Chamberlain et al., 2016), les ambiguïtés référentielles sont annotées comme des alternatives avec des scores correspondant aux avis des annotateurs. Pour le polonais, Ogrodniczuk et al. (2015) distinguent les relations de coréférence identique et quasi-identique. Le projet ANCOR (Muzerelle et al.,

2013), qui traite du français oral, préconise de caractériser les relations entre les mentions selon 5 types : directe, indirecte, pronominale, associative et associative pronominale. Ces corpus distinguent donc différents types de relations entre les mentions mais toujours de manière stricte.

Annoter les phénomènes de coréférence non stricte et floue permettrait d'obtenir des chaînes de coréférence qui reflètent mieux le texte en gardant l'information que certaines expressions semblent correspondre à une chaîne sans coréférer à ses maillons de manière stricte. C'est pourquoi nous proposons un schéma d'annotation répertoriant trois catégories de coréférence qui prend en compte la coréférence non stricte. La première catégorie correspond à la coréférence exacte : lorsque les mentions réfèrent exactement et sans aucun doute au même référent, comme c'est le cas dans le manuel d'annotation du projet Democrat (Landragin, 2016) par exemple. La seconde catégorie correspond à la coréférence inclusive, elle comprend les cas où un référent en inclut complètement un autre, de manière stricte ou de manière floue. La troisième catégorie correspond à la coréférence intersective. Il s'agit des cas où la coréférence n'a lieu que sur l'intersection de deux référents, de manière stricte ou de manière floue.

Le schéma d'annotation proposé s'inspire du schéma proposé par le projet Democrat, bien que les relations ne soient pas annotées. Il correspond au modèle d'annotation de type Unités-Relations-Schémas (URS), développé à l'origine dans le logiciel Glozz (Widlöcher & Mathet, 2009) et implémenté dans le logiciel Analec (Landragin, 2012) puis par extension dans le logiciel TXM (Heiden, 2010). L'annotation du type de coréférence se fait au niveau des relations pour chaque couple de mentions. Chaque catégorie est représentée par un trait : exacte, inclusive ou intersective. Pour les catégories de coréférence inclusive et intersective, il est possible de choisir entre les deux propriétés : stricte et floue. Ce schéma est en cours de test sur une partie du projet Democrat dans le but d'effectuer une comparaison avec un texte déjà annoté en coréférence stricte. Ce schéma d'annotation sera décrit dans un manuel d'annotation qui pourra répertorier les critères caractérisant chaque catégorie ainsi que des exemples issus du corpus. Pour valider ce schéma, des tests avec de la double annotation sont aussi prévus.

Références bibliographiques

Chamberlain, J., Poesio, M., & Kruschwitz, U. (2016). Phrase Detectives Corpus 1.0 Crowdsourced Anaphoric Coreference. *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 2039-2046. Portorož, Slovenia.

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

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., & Weischedel, R. (2004). The Automatic Content Extraction (ACE) Program –Tasks, Data, and Evaluation. *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, 837-840. Lisbon, Portugal.

Ghaddar, A., & Langlais, P. (2016). *WikiCoref : An English Coreference-annotated Corpus of Wikipedia Articles.* Présenté à Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia.

Heiden, S., Magué, J.-P., & Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. *Proc. of 10th International Conference on the Statistical Analysis of Textual Data*, 2, 1021-1032. Edizioni Universitarie di Lettere Economia Diritto, Roma, Italy.

Landragin, F. (2016). Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). *Bulletin de l'AFIA*, (92), 11-15.

Landragin, F., Poibeau, T., & Victorri, B. (2012). ANALEC : a New Tool for the Dynamic Annotation of Textual Data. European Language Resources Association (ELRA). *International Conference on Language Resources and Evaluation*, 357-362. Istanbul, Turkey.

Muzerelle, J., Lefeuvre, A., Antoine, J.-Y., Schang, E., Maurel, D., Villaneau, J., & Eshkol, I. (2011). ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In ATALA (Éd.), *20e conférence sur le Traitement Automatique des Langues Naturelles* (p. 555-563). Les Sable d'Olonne, France : ATALA.

Ogrodniczuk, M., Glowinska, K., Kopec, M., Savary, A., & Zawislawska, M. (2014). *Coreference : Annotation, Resolution and Evaluation in Polish.* Walter de Gruyter GmbH & Co KG.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., & Xue, N. (2011). CoNLL-2011 Shared Task : Modeling Unrestricted Coreference in OntoNotes. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning : Shared Task*, 1-27. Portland, Oregon, USA.

Recasens, M., Hovy, E., & Martí, M. A. (2011). Identity, non-identity, and near-identity : Addressing the complexity of coreference. *Lingua*, 121(6), 1138–1152.

Widlöcher, A., & Mathet, Y. (2009). La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. *Actes de la 16e Conférence Traitement Automatique des Langues Naturelle, session posters*, 10. Senlis, France.

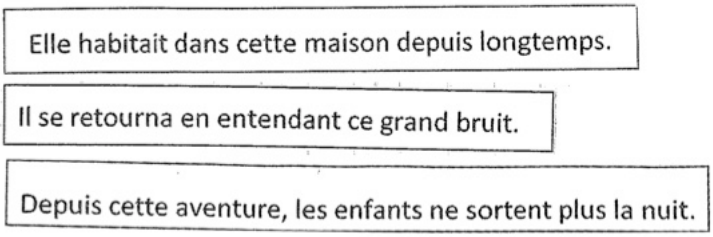
ResolCo un corpus de manuscrits d'élèves et d'étudiants pour l'étude de la cohérence

Lydia-Mai Ho-Dac , Silvia Federzoni , Myriam Bras , Josette Rebeyrolle et Claudine Garcia-Debanc

CLLE, Université Toulouse Jean Jaurès

lydia-mai.ho-dac@univ-tlse2.fr, silvia.ferderzoni@univ-tlse2.fr, myriam.bras@univ-tlse2.fr, josette.rebeyrolle@univ-tlse2.fr, claudine.garcia-debanc@univ-tlse2.fr

La ressource ResolCo[1] est un corpus constitué de transcriptions de manuscrits d' élèves et d'étudiants enrichi par des annotations concernant les traces du processus d'écriture, les variantes orthographiques observées et certaines structures discursives. L'originalité de la ressource ResolCo réside dans son protocole de collecte et principalement dans sa consigne qui est la suivante : *Racontez une histoire dans laquelle vous insérerez, séparément et dans l'ordre donné, les trois phrases suivantes (découpez et collez les bandelettes dans votre texte)*



Cette consigne a été imaginée pour provoquer la mise en œuvre de stratégies de Résolution de problèmes de Cohérence (Charolles 1994, Garcia-Debanc et al. 2017). Le corpus ResolCo fournit ainsi un terrain privilégié pour l'étude de l'organisation du discours et des indices de cohésion à différents âges d'acquisition de la langue écrite. Les stratégies mis en jeux sont principalement :

- les stratégies utilisées pour introduire des référents de type variés (humains –*Elle, Il, les enfants* ; inanimés –*cette maison, ce grand bruit* ; événementiel –*cette aventure*) et gérer la compétition et l'interférence entre les continuités référentielles (cf. Ariel 1990 et Givon 1983) ;
- des stratégies de planification du discours (amorçage de la phrase-fermoir (cf. Marandin 1986) et gestion de l'anaphore résumante – *cette aventure*) ;
- des stratégies de gestion des temps verbaux, chaque phrase présentant un temps du récit différent ;
- la production d'un texte de type narratif sans contrainte de genre.

Les productions écrites ont été recueillies dans différentes écoles primaires, collèges et universités des régions Occitanie et Île de France. Les points de collectes retenus permettent de

1. La ressource ResolCo fait partie du projet ANR É :Calm, Écritures scolaires : Corpus, Analyses Linguistiques, Modélisations didactiques. Voir `http://e-calm.huma-num.fr/`

disposer d'un nombre comparable de copies par niveau et pour représenter différents milieux scolaires (urbain, rural, ZEP). Toutes les données récoltées seront mises à disposition de la communauté sous licence Creative Commons By-NC-SA 3.0 (Paternité, usage non commercial, partage à l'identique).

La première étape de constitution du corpus consiste en la transcription et l'annotation des traces écrites avec pour objectif de reproduire le plus fidèlement possible la copie originale et sa mise en page. Pour ce faire, les transcriptions sont accompagnées d'annotations indiquant d'éventuelles ratures, la présence des dessins des élèves, etc. Le format choisi pour la transcription est le format XML et la norme TEI-P5 qui fournit à la fois un modèle pour l'encodage des métadonnées et de nombreux éléments dédiés à l'annotation des traces écrites.

Les métadonnées retenues concernent le contexte de production de la copie (établissement, niveau scolaire, année scolaire, consigne d'écriture, etc.). La norme TEI-P5 permet également d'indiquer les étapes de digitalisation des données. Concernant le corps de texte, en plus de l'indication de l'emplacement des bandelettes, les annotations signalent des phénomènes de mise en page et d'écriture manuscrite.

Les principales conventions de codage TEI-P5 des traces écrites adoptées sont les suivantes : chaque ligne sur la copie est encodée par l'élément <lb>, les paragraphes correspondent à l'élément <p>, tout trace de texte révisé (insertion et/ou suppression par rature ou effacement visible) est renseignée par l'élément <mod> et visualisée au moyen d'un texte en exposant et/ou barré, toute portion de texte illisible ou à la transcription incertaine est indiquée (<gap> ou <unclear>) est affichée sous forme d'italiques ou de séquences de "xx".

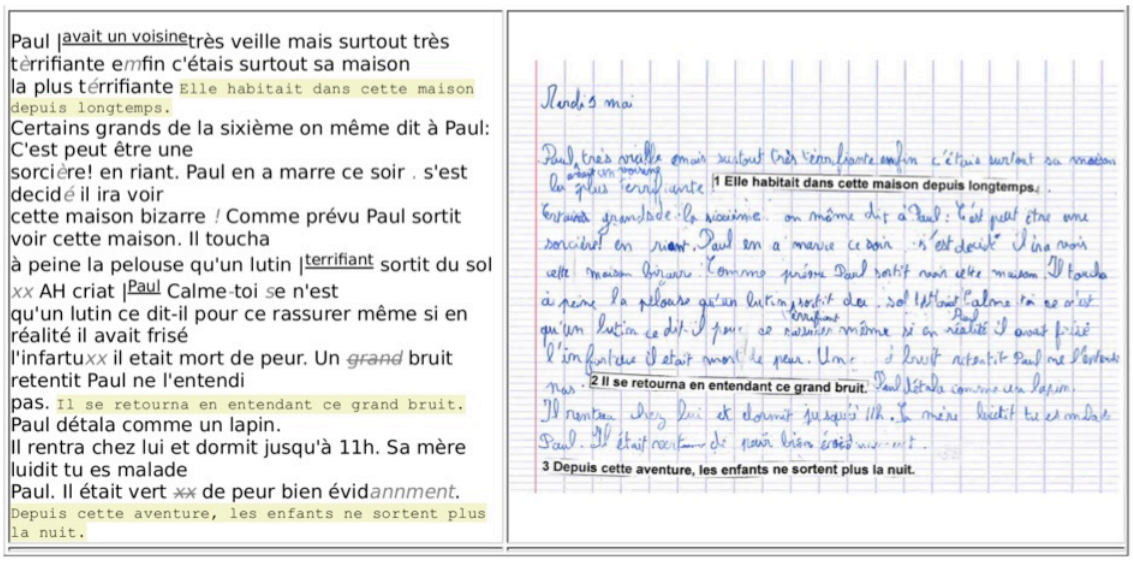


FIG. 1 : Transcription des traces d'écriture et scan d'une copie

Afin de s'assurer de la qualité des données, toutes les transcriptions ont été vérifiées par un annotateur différent du transcripteur. Cette phase de vérification s'est appuyée sur un affichage simultané du scan de la copie originale et d'une visualisation du texte transcrit obtenue par transformation XSLT depuis le fichier XML, comme l'illustre la figure 1.

La transcription XML est ensuite converti au format Glozz (Widlöcher & Mathet 2009) pour permettre l'annotation des variantes orthographiques et la génération d'une version normée permettant l'application d'outils du TAL et l'annotation des structures discursives. Le tableau 1 donne un aperçu quantitatif de l'état actuel du corpus ResolCo en indiquant pour chaque niveau : le nombre de textes transcrits, de ratures, de textes normés et d'erreurs d'orthographe.

	ratures	textes transcrits	ratures/texte	corrections	textes normés	corrections/texte
total	**1590**	**350**	**5**	**1336**	**132**	**10**
CE2	97	36	3	199	13	15
CM1	214	39	5	29	25	1
CM2	299	94	3	371	35	11
6EME	396	85	5	687	29	24
4EME	204	47	4	0	0	na
3EME	276	36	8	50	17	3
Master	91	13	7	0	13	0

TAB. 1 : aperçu quantitatif de l'état actuel de la ressource ResolCo

La version normée est également associée à un étiquetage des catégories morphosyntaxiques et des relations syntaxiques produit par l'outil Talismane (Urieli 2013). La constitution d'un treebank par correction manuelle des analyses proposées par Talismane a permis une évaluation sur un échantillon de 13 220 tokens dont 11 706 mots (hors ponctuations).

Les résultats obtenus sont tout à fait acceptables avec une exactitude de 95,7 pour l'attribution des catégories morphosyntaxiques (11 203 token correctement étiquetés sur 11 706) et une efficacité au niveau de l'analyse des relations syntaxiques (i.e. détecter le gouverneur et le type de relation entre chaque token) de 97,5 pour la détection du bon gouverneur et de 90,7 pour la caractérisation des relations.

Cette ressource permet un ensemble d'analyses fournissant des données originales sur l'évolution des compétences scripturales liées notamment à l'orthographe, la syntaxe et le discours au fil de l'acquisition de l'écriture à l'école.

Références bibliographiques

Ariel, M. (1990). *Accessing noun phrase antecedents.* Routledge : London

Charolles, M. (1994). Cohésion, cohérence et pertinence du discours. *Travaux de Linguistique*, 29, 125-151

Garcia-Debanc, C., Ho-Dac, L.-M., Bras, M., Rebeyrolle, J. (2017). Vers l'annotation discursive de textes d'élèves. *Corpus*, 16.

Givón, T. (1983). *Topic continuity in discourse : an introduction. In T. Givon (ed) Topic continuity in discourse : a quantitative cross-language study.* John Benjamins : Amsterdam/Philadelphia, pp. 1-42

Marandin, J. M. (1986). *Ce est un autre.* L'interprétation anaphorique du syntagme démonstratif. *Langages*, 81, pp. 75-89.

Urieli, A. (2013). *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit.* Thèse de doctorat, Université Toulouse le Mirail-Toulouse II.

Widlöcher, A., & Mathet, Y. (2009). La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. In *Actes de la 16e Conférence Traitement Automatique des Langues Naturelles (TALN'09).*

Session 3.B.
Lexique et collocations

A preliminary sketch of ”Brexit” and ”tourism”: A corpus-driven analysis of their relationship as deployed by the press.

Camino Rea Rizzo
Universidad Politécnica de Cartagena
Camino.rea@upct.es

Abstract

On 23rd June 2016 a crucial referendum for Europe was held: the United Kingdom voted on its permanence or withdrawal from the European Union. The consequent leave result led to the onset of Brexit. Research on this phenomenon throughout its short span of existence has been carried out from multiple angles. Among them, political studies were initially conducted to ascertain the reasons why a leave result won the referendum and the factors on which public support for Brexit would depend (Goodwin and Heath, 2016); social studies pursued the same objective adopting as object of analysis the public tweets posted on the event taking advantage of digital means (Maynard et al., 2017); subsequently, the focus of study was directed towards the influence of the media on the leave result and how such bias was conveyed by the press (Seaton, 2016). In turn, the multifaceted mode of analysis enabled by language studies covers different means of communication on the subject, such as social media platforms (Griebel and Heinrich, 2017) and in particular, news articles (Ballmann, 2017). The latest volume of works in the field at the time of writing (March 2019) presents "*the first comprehensive exploration of discourses surrounding the UK's departure from the EU and as such step towards understanding the reasons for, and processes of, Brexit*" (Koller, Kopf and Miglbauer, 2019:1). The present work provides further insight into the press discourse on one of the areas that would be significantly affected by the withdrawal of the United Kingdom from the European Union, namely, the tourism sector and hospitality industry (Rhodes and Ward, 2016).

The objective of this work is to identify how the online press echoes the relationship between the consequences of Brexit on tourism, which originally tended to be quite pessimistic. For that purpose, a linguistic corpus of online press news is designed and compiled ad hoc – following the principles of Corpus Linguistics (Sinclair, 1991; EAGLES, 1996; Rea, 2010), in order to point out the lexical choice of the press when reporting on the issue. The so-called News Corpus is a collection of online newspaper written news delivered from the immediate aftermath of the referendum to the beginning of 2018. The interval of data collection leads to an unavoidable temporary overview of the situation, since the Brexit process is still under development. The pieces of news deal with the specific subject of tourism and Brexit from British and non-British newspapers addressed to the general readership being. The samples were collected in accordance with the BOAI definition of open access, that is, they were freely available on the public internet. The size of the corpus reaches 201,108 tokens (9,183 types; 4.7 type/token ratio).

The corpus is processed through Wordsmith software (Scott, 1996) using the tools available and the parameters which determine keywords. In order to execute the Keyword tool, the general language corpus LACELL (`https://www.um.es/grupos/grupo-lacell/proyectos.php#dos`) has been used as a reference corpus. The analysis shows that Brexit ranks as the second highest keyword (index: 11,661) and *tourism* as the fourth one (index: 9,535). Due to their high keyness and being the focus of analysis, their surrounding co-text is examined in order to describe the syntagmatic relations established, as shown by the collocates and word clusters that both generate (See a sample of their top 3-word clusters Table 1). Such results combined with a closer qualitative assessment would reveal the impact and consequences of the Brexit in the tourist sector.

***Tourism* clusters**	**Freq.**	***Brexit* clusters**	**Freq.**
THE TOURISM INDUSTRY	90	THE BREXIT VOTE	95
TRAVEL AND TOURISM	80	IMPACT OF BREXIT	76
OF BREXIT ON	32	OF BREXIT ON	59
THE IRISH TOURISM	26	THE IMPACT OF	47
TOURISM IN THE	26	OF THE BREXIT	46
THE TOURISM SECTOR	26	A POST BREXIT	27
IRISH TOURISM INDUSTRY	26	AS A RESULT	27
OF THE TOURISM	24	SINCE THE BREXIT	25
IMPACT OF BREXIT	23	RESULT OF BREXIT	24
AND TOURISM INDUSTRY	23	A RESULT OF	24
THE UK TOURISM	22	THE BREXIT REFERENDUM	19
S TOURISM INDUSTRY	21	FOLLOWING THE BREXIT	17
FOR IRISH TOURISM	21	AFTER THE BREXIT	16
IN THE TOURISM	21	EFFECTS OF BREXIT	16
TOURISM INDUSTRY IS	20	BREXIT ON TOURISM	16
THE UK S	19	TRAVEL AND TOURISM	15
IN THE UK	19	BREXIT WILL HAVE	15
TOURISM INDUSTRY AND	18	BREXIT MEAN FOR	15
FOR THE TOURISM	18	THE WAKE OF	15
AND TOURISM SECTOR	18	THE EFFECTS OF	15
UK TOURISM INDUSTRY	17		
TOURISM FROM THE	17		
BREXIT ON TOURISM	16		
TOURISM TO THE	15		
THE TOURISM ALLIANCE	15		
IMPACT ON TOURISM	14		
TO IRISH TOURISM	14		
TOURISM IS ONE	14		
FOR INBOUND TOURISM	13		
IRELAND S TOURISM	13		
TOURISM INDUSTRY THE	13		
THE TRAVEL AND	13		

Table 1: *Tourism* and *Brexit* top 3-word clusters.

In addition, when the keyword list is studied as a whole, it is possible to identify three main groups of words depending on the field of activity to which they belong: politics, economy and tourism. Moreover, a first observation of the top 150 keywords seems to provide the answer to the well-known 5 Ws (Who, What, Where, When and Why) that a news story must cover in order to tell the readership what they need to know in the most straightforward and shortest possible

way.

After accomplishing the study, the data obtained lead to conclude that the pessimistic forecast for the tourism sector following the departure of the UK from the EU have not been met (so far); as a matter of fact, the opposite effect has occurred. Notwithstanding, a pervasive sense of uncertainty and overriding concern over inbound and outbound traveller's future is still perceived.

References

Ballmann, Katja (2017). *Brexit in the news: – frames and discourse in the transnational media representation of Brexit.* Stockholm University. `http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1107235&dswid=9217` (Retrieved 12/03/2019)

EAGLES (1996). Preliminary Recommendations on Corpus Typology. Expert Advisory Group on Language Engineering. EAG-TCWG-CTYP/P.

Goodwin, Matthew J. and Oliver Heath (2016). The 2016 Referendum, Brexit and the Left Behind: An Aggregate-level Analysis of the Result. *The Political Quarterly*, Vol. 87, No. 3, (323-332).

Griebel, Tim and Philipp Heinrich (2017). The Cultural Political Economy of Brexit in the Age of Austerity. A Corpus-Assisted Critical Realist Multimedia Discourse Analysis. Paper presented at the *IPSA Conference "Political Science in the Digital Age"* Hannover, December 4-6, 2017. Work in Progress.

Koller, Veronika, Susanne Kopf, Marlene Miglbauer Eds. (2019). *Discourses of Brexit*. Routledge.

Maynard, Diana, Ian Roberts, Mark A. Greenwood, Dominic Rout, Kalina Bontcheva (2017). A framework for real-time semantic social media analysis. *Journal of Web Semantics,* Volume 44, (75-88).

Rea Rizzo, Camino (2010). Getting on with corpus compilation: from theory to practice. *English for Specific Purposes World*. Issue 1 (27), Vol. 9.

Rhodes, Chris and Ward, Matthew (2016). *Potential effect of the UK leaving the EU on UK tourism.* Briefing Papers. The House of Commons Library.

Seaton, Jean (2016). Brexit and the Media. *The Political Quarterly,* Vol. 87, No.3, (333-337).

Scott, M (2010). Wordsmith Tools software. Version: 5.0.

Sinclair, J. (1991). *Corpus, Concordance and Collocation.* Oxford: Oxford University Press.

Analyse collocationnelle des verbes « *provide* » et « *realizar* » dans le corpus de rapports RSE « GRIC » : une mise en évidence d'éléments de contingence culturelle en contexte normalisé

Emmanuelle Pensec
LIDILE, Université de Rennes 2
LEGO, Université de Bretagne Sud
Emmanuelle.pensec@univ-rennes1.fr

1 Introduction

Ce papier examine le discours déployé par les organisations pour décrire les actions mises en œuvre dans le cadre des politiques de Responsabilité Sociale des Entreprises (RSE, dorénavant). Il se concentre sur l'analyse de la description de la relation entre l'organisation et ses parties prenantes dans deux contextes culturels différents. En considérant l'impact du référentiel de normalisation du reporting RSE de la Global Reporting Initiative comme facteur d'homogénéisation du discours dans les rapports RSE, la question de recherche proposée est la suivante : comment les choix discursifs des organisations dans les rapports RSE normalisés traduisent-ils une forme de contingence culturelle ? Pour répondre à cette question, j'ai comparé des données textuelles extraites d'un corpus de rapports RSE normalisés préalablement construit par mes soins, nommé GRIC, afin de constituer deux corpus géographiques, reflétant les publications de rapports RSE d'organisations étatsuniennes et espagnoles. Cette analyse inductive combine une méthode quantitative et qualitative. Elle s'inscrit dans la linguistique de corpus contextualiste dans la lignée de Firth (Firth, 1957) et de Sinclair (Sinclair, 2005) pour observer la langue telle qu'elle est réellement utilisée et ainsi déterminer la manière dont l'organisation se met en scène dans sa relation avec ses parties prenantes. Les résultats de l'analyse montrent un positionnement foncièrement distinct selon la nationalité de l'organisation.

2 Corpus et méthodologie

2.1 Corpus

Afin d'identifier le positionnement de l'organisation dans sa relation avec ses parties prenantes, j'ai comparé un corpus de rapports RSE publiés par des organisations étatsuniennes composé de 84 rapports RSE, soit 1 722 766 *tokens*, à un corpus de rapports RSE publiés par des organisations espagnoles constitué de 33 rapports RSE, soit 1 312 902 *tokens*. Ces deux corpus ont été élaborés dans le même contexte, c'est-à-dire : d'une part, sur la période 2006-2011, correspondant à une crise de légitimité pour les organisations en raison de la crise économique des *subprimes* ; d'autre part, selon le référentiel de rédaction de la Global Reporting Initiative, norme de reporting la plus utilisée à l'échelle internationale depuis le début des années 2000.

Ainsi, les données textuelles analysées sont comparables et représentatives du phénomène managérial et discursif que constitue le rapport RSE (Biber et al., 1998 ; Williams, 2002 ; Sinclair, 2005). L'ensemble des organisations retenues pour l'analyse sont des entreprises privées issues

des secteurs d'activités les plus publiants car considérés « à risque », à savoir : les services financiers, l'agroalimentaire, l'énergie et les services énergétiques, la chimie et le secteur de la mine.

2.2 Méthodologie

Afin d'identifier les régularités discursives caractéristiques des deux corpus, j'ai adopté la Théorie des Réseaux Collocationnels (Williams, 1998) ainsi que la *Corpus Pattern Analysis* (Hanks, 2004, 2013). Cette approche combinée permet de sélectionner les items lexicaux prototypiques d'un corpus thématique afin d'observer les patrons associés à ces items et déterminer un cadre de référence thématique. Il est alors possible d'identifier le lexique spécialisé de la thématique traitée ainsi que ses collocations et patrons les plus significatifs. Les régularités ainsi détectées dans les corpus permettent de mettre en évidence des paradigmes collocationnels (Williams, DeCesaris, et Alonso Campo, 2017) qui contribuent à la définition des arguments les plus caractéristiques de chaque discours. Dans l'optique de la question de recherche soulevée au début de ce papier, l'objectif de mon analyse consiste à déterminer la manière dont les ressources textuelles sont exploitées par les organisations pour s'adapter aux destinataires du discours RSE dans un rapport normalisé. L'analyse collocationnelle des deux corpus est réalisée avec l'aide de l'outil lexicographique *Sketch Engine* (Kilgarriff, 2004 ; Kilgarriff et al., 2014). A partir de listes de lemmes les plus fréquents, je constate des choix lexicaux différents selon les corpus. Ainsi, deux verbes sont retenus pour l'analyse en raison de leur représentativité dans chaque ensemble textuel : « *realizar* » pour le corpus espagnol et « *provide* » pour le corpus étatsunien dont je détermine les collocats les plus significatifs pour construire deux réseaux collocationnels présentés figures 1 et 2.

L'analyse de chacun de ces réseaux permet effectivement de déterminer des régularités et des différences spécifiques à chaque environnement discursif.

3 Résultats

Le réseau collocationnel du verbe « *provide* » (figure 2) traduit le positionnement de l'organisation étatsunienne dans sa relation avec ses parties prenantes. L'analyse permet de faire émerger un emploi prototypique de deux patrons :

[we] + [also] + [provide] + [information / service / opportunity / training / support]
[we] + [provide] + [customer / employee] + [with]

Le réseau met en lumière une relation contractuelle et marchande assumée avec les collaborateurs et clients, plaçant l'organisation dans un rôle de fournisseur, tel le maillon d'une chaîne au même titre que les autres acteurs en présence. L'emploi du pronom sujet « *we* » souligne un engagement dans les actions menées, renforcées par l'emploi régulier du modifiant « *also* » qui souligne l'image active de l'organisation. Les actions menées sont concrètes et largement

FIG. 1 : Réseau collocationnel du verbe « *realizar* » dans le corpus espagnol

décrites par les collocats « *information* », « *service* », « *opportunity* », « *training* » et « *support* ». L'organisation s'inscrit à la fois dans une démarche économique et collective où l'action de chacun contribue à la réussite de tous. Le discours RSE dans le corpus étatsuninen est un discours de l'action. Le réseau collocationnel du verbe « *realizar* » (figure 1), quant à lui, présente une image de l'organisation dans un rôle de supervision. Les collocats du verbe ne sont pas orientés vers le sujet de l'action mais plutôt vers la description de ses réalisations. En ce sens, les collocats du verbe « *realizar* » relèvent majoritairement de la catégorie grammaticale « complément » et fournissent deux types d'informations relevant d'une forme de contingence. Tout d'abord, le réseau collocationnel met en évidence l'emploi de l'item « *actividad* » pour décrire l'activité de l'organisation, là où le corpus étatsunien privilégie l'item « *business* ». Le mot « *actividad* » est dénué de connotation marchande contrairement au mot « *business* » et met en valeur le savoir-faire. Ensuite, le réseau met également en évidence un discours insistant sur l'évaluation et la vérification (« *encuesta de satisfacción* », « *estudio de satisfacción* ») ainsi que sur la conformité des actions réalisées et des résultats obtenus (« *evaluación* », « *auditoría* », « *seguiminento de riesgo* », etc.). L'analyse des collocats de « *realizar* » montre une organisation davantage orientée vers un rôle d'analyse et de réflexion. Les résultats de l'analyse collocationnelle mettent ainsi en évidence un discours qui traduit deux approches communicationnelles de la relation entre l'organisation et ses parties prenantes. D'une part, l'organisation étatsunienne s'inscrit dans une approche utilitariste et explicite de la RSE (Matten et Moon, 2008) par un dis-

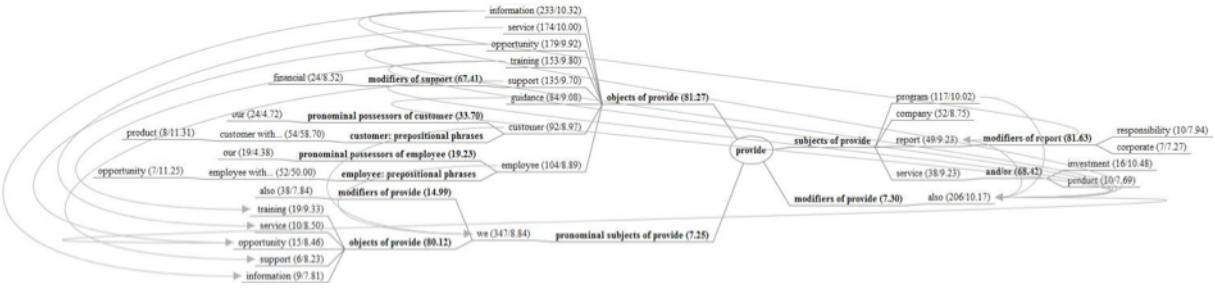


FIG. 2 : Réseau collocationnel du verbe « *provide* » dans le corpus étatsunien

cours orienté sur les actions mises en œuvre en matière de RSE, notamment grâce à de nombreux éléments concrets relatifs à ses pratiques. D'autre part, l'organisation espagnole relève davantage d'une approche macro sociale et implicite (Matten et Moon, 2008) par un discours s'inscrivant dans une réflexion et une conceptualisation des enjeux relatifs à la RSE.

Références bibliographiques

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics : Investigating Language Structure and Use* (Cambridge University Press). Cambridge.

Firth, J. R. (1957). *Papers of Linguistics 1939-1951* (OUP). London.

Hanks, P. (2004). Corpus Pattern Analysis. In *Euralex 2004 Proceedings*.

Hanks, P. (2013). *Lexical Analysis : Norms and Exploitations* (The M.I.T. Press). Cambridge.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In *Proceedings of Euralex* (p. 105-116). Lorient, France.

Kilgarriff, A., Baisa, V., Busta, J., Jakubiecek, M., Kovai, V., Michelfeit, J., & Suchomel, V. (2014). The Sketch Engine : ten years on. *Lexicography*, 1(1), 7-36.

Matten, D., & Moon, J. (2008). Implicit and explicit CSR : A conceptual framework for a comparative understanding of corporate social responsibility. *Academy of Management Review*, 33(2), 404-424.

Sinclair, J. (2005). Corpus and Text. In *Developing Linguistic Corpora : A Guide to Good Practice* (AHDS Literature, Languages and Linguistics). Oxford : M. Wyne.

Williams, G. (1998). Collocational networks : Interlocking patterns of lexis in a corpus of plant biology. *International Journal of Corpus Linguistics*, 3(1), 151-171.

Williams, G. (2002). In search of representativity in specialised corpora : Categorisation through collocation. *International Journal of Corpus Linguistics*, 7(1), 43-64.

Williams, G., Alonso Campo, A., & DeCesaris, J. (2017) Studying lexical meaning in context : From collocation to collocational networks and resonance. In *Collocations and Other Lexical Combinations in Spanish* (Routledge). London : S. Torner Castells & E. Bernal Gallen.

Session 4.A.
Construction de corpus

Building an informal and conversational corpus. Design, field work and annotation in a spoken corpus for Catalan language

Andreu Sentí
Departament de Filologia Catalana, Universitat de València
andreu.senti@uv.es

1 Introduction

The design of a spoken corpus entails several methodological, technological and linguistic issues (cf. Love et al. 2017; Lüdeling & Kytö 2008; Wichmann 2008). Unlike written corpora, spoken language presents, amongst other, two challenges for corpus linguistics. First, the primary source always has to be the acoustic sound of the voice, but also a written code is required which allows to process the data and retrieve the information in a proper way for future studies. Secondly, even though pre-existing data can be collected, typically it is not the case and it is necessary to create the situation in which different people speak spontaneously while they are being recorded. In this presentation I aim to deal with these two issues in relation to the design of an informal spoken corpus for Catalan language.

2 Corpus and methodology

2.1 Corpus

Spoken Catalan language has a deficit of corpora, especially for conversational, colloquial and dialectal speech (cf. Beltran *et al.* in press). The *Parlars* corpus is an ongoing project that aims:

a. To document the least corrupted informal Catalan (cf. Beltran & Segura-Llopes 2017), before it disappears in the face of the pressure exerted by the standard form of the language and, above all, Spanish, which is diluting the language at a great rate throughout the areas where it is spoken (Segura 2003).

b. To provide researchers with suitable materials to carry out descriptive and analytical studies of the linguistic variation in Catalan, especially functional (colloquial) and dialectal.

c. To test the hypothesis that languages function as conventionalized structures, as described by Cognitive Construction Grammar (Goldberg 2003; Taylor 2012; Hilpert 2014).

2.2 Méthodologie

In the frame of the design of the *Parlars* corpus, this presentation will focus on two methodological issues and one application of Corpus linguistics: 1) data collection methodology; 2) data processing (transcription and annotation schema); 3) the linguistic analysis of some constructions according to preliminary results (evidential and modal Catalan constructions mainly).

In the first place, I will talk about the speakers. The current project covers only the Valencian dialect, i.e. all the 22 Catalan-speaking districts in the Valencian Country. In the *Parlars* corpus each district is represented by one, two or three villages depending on its geographic and linguistic characteristics: for example, it is necessary to represent urban and rural models, coastal and inland models, and some specific cases. Only old people (65-95 years) are selected. The ideal speaker has little structure and involves father and mother or grandfather and grandmother from the same locale in order to find a representative speaker of the local dialect. Most of our speakers share this feature.

Then, I will explain the different techniques for eliciting/collecting data. The types of speech (or interactions) that have been elicited are secret and spontaneous conversational speech, non-secret (semi)spontaneous conversation (guided by an interviewer), personal history narratives and computer-mediated informal speech (mainly voice messages in a Whatsapp dialogue). The differences and similarities between spontaneous conversation (prototypical, according to Briz 2010) and non-secret conversation will be analysed. The colloquial features attested in each kind of interaction will be discussed in order to distinguish actual prototypical spontaneous speech from other types of speech. Besides, the results of other methodological strategies to elicit dialogic speech will be presented as well as, for example, the absence of a researcher, the presence/absence of a local team member who is family to the informants, the introduction of different topics for the conversation, etc. (cf. Basanta 2018).

The second important methodological issue is data processing. The transcription and annotation work are done with the open and free tool source ELAN (Wittenburg *et al.* 2006), widely used for the transcription of multimedia documents. Among other advantages, ELAN allows to define multiple tiers: transcription, tokenisation, part of speech, lemma, and other annotations.

The transcription schema (Beltran *et al.* 2019) is near to the signal of speech regarding morphological, syntactic and lexical peculiarities, but at the same time it is a wider transcription avoiding dialectal phonetics. Therefore, the result is an orthographic text that facilitates the (semi)automatic annotation. Unlike previous transcription schemata (Payrató & Alturo 2002; Briz 2002; Hidalgo & Sanmartín 2005; Bladas 2009), this model facilitates the (semi)automatic lemmatisation work with Apertium (Forcada *et al.* 2011) (tokenisation, lemmatisation, morpho-syntactic annotation and alignment) (cf. Ide & Pustejovsky 2017). Although Apertium provides a tool for Catalan language, it is not prepared for dialectal and spoken language, for that reason new rules and expanded paradigms are required to be done manually.

3 Results

Finally, I will show some results from the fieldwork and the conversations registered so far within the project of the aforementioned corpus. With the intention of approaching the analysis of constructions in the spoken language (cf. Taylor 2012; Hilpert 2014, 2018), I will deal with Catalan modal verbs and other constructions and verbs that have assumed evidential or epistemic

values. Some of these constructions can only be found in spoken corpus, such as *diu que* 'it is said that', because it is restricted to informal speech (Antolí & Sentí in press):

a. **ho havien de** prohibir (Parlars corpus, Benissa)
'that **should** be forbidden'

b. **tenien que** treballar (Parlars corpus, Benissa)
'they **had** to work'

c. ara no sé si en **quedarà** una o dos, si en queden (Parlars corpus, Benissa)
'and now I don't know if there **must be** a shop or two opened in town'

d. allò **diu que** està molt vell (Parlars corpus, Benissa)
'**it is said that** that is very old'

e. **diu que** si ja s'ho deixen (Parlars corpus, Benissa)
'**it is said that** they ara going to abandon'

f. **Trobe** que va ser… (Parlars corpus, Benissa)
'**I think** it was…'

In conclusion, my presentation will try to shed some light on one of the most challenging research questions when it comes to spoken language within the context of corpus linguistics: What can an informal spoken corpus tell us about spoken language?

References

Antolí, Jordi & Sentí, Andreu (in press). Evidentiality in spoken Catalan. The evidential construction *diu que. Anuari de Filologia. Estudis de lingüística*, 10. Barcelona, Universitat de Barcelona.

Basanta, Noemi (2018). "As formas cambiaron porque o mundo cambiou": construción discursiva e interseccional de identidades de xénero e sexualidade en conversas sobre ligar. PhD dissertation. Universidade de Santiago de Compostela.

Beltran, Vicent & Segura-Llopes, Carles (2017). *Els parlars valencians*. València, PUV.

Beltran, Vicent; Esplà, Miquel; Guardiola, M. Isabel; Montserrat, Sandra; Segura, Carles; Sentí, Andreu (2019). *Criteris per a la transcripció del corpus Parlars. Segona versió*. València, Universitat de València. Roderic. <http://roderic.uv.es/handle/10550/71244>

Beltran, Vicent; Esplà, Miquel; Guardiola, M. Isabel; Montserrat, Sandra; Segura, Carles; Sentí, Andreu (in press). El corpus parlars. *Zeitschrift für Katalanistik.*

Bladas, Òscar (2009). *Manual de transcripció del discurs oral. Materials de treball.* Universitat de Barcelona.

Briz Gómez, Antonio y Grupo Val.Es.Co. (2002). Corpus de conversaciones coloquiales, *Anejo de la revista Oralia*. Madrid, Arco-Libros.

Briz, Antonio (2010). El registro como centro de la variedad situacional. Esbozo de la propuesta del grupo Val.Es.Co. sobre las variedades diafásicas. In Fonte, I.; Rodríguez Alfano, L. (ed.): *Perspectivas dialógicas en estudios del lenguaje*. México, Universidad Autónoma de Nuevo León, 21-56.

ELAN (Version 5.2) [Computer software]. (2018, April 04). Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from `https://tla.mpi.nl/tools/tla-tools/elan/`

Forcada, Mikel L. & Ginestí-Rosell, Mireia & Nordfalk, Jacob & O'Regan, Jim & Ortiz-Rojas, Sergio & Pérez-Ortiz, Juan Antonio & Sánchez-Martínez, Felipe & Ramírez-Sánchez, Gema & Tyers, Francis M. (2011). Apertium: a free/open-source platform for rule-based *machine translation*. Machine translation 25.2 (2011): 127-144.

Hidalgo, A. & J. Sanmartín, (2005). Los sistemas de trascripción de la lengua hablad, *Oralia*, 8, 13-36.

Hilpert, Martin (2014). *Construction Grammar and its Application to English*. Edimburgh, Edimburgh University Press.

Hilpert, Martin (2018). Construction Grammar and the analysis of spoken language. Presentation at LingCor2018. 1st International Workshop on Spoken Corpus Linguistics. València: Universitat de València.

Ide, Nancy & James Pustejovsky (eds.) (2017). *Handbook of Linguistic Annotation*. Dordrecht, Springer.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014. *International Journal of Corpus Linguistics.*

Lüdeling, A., & Kytö, M. (eds.) (2008) *Corpus Linguistics: An International Handbook*. Berlin: Walter de Gruyter.

Payrató, Lluís & Núria Alturo (ed.) (2002). *Corpus oral de conversa col·loquial. Materials de treball*. Barcelona, Publicacions de la Universitat de Barcelona.

Segura, Carles (2003). *Variació dialectal i estandardització al Baix Vinalopó*. Alacant / Barcelona, Institut Interuniversitari de Filologia Valenciana / Publicacions de l'Abadia de Montserrat.

Taylor, John R. (2012). *The mental corpus*. Oxford: Oxford University Press.

Wichmann, A. (2008). Speech corpora and spoken corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook*. Berlin, Walter de Gruyter, 187–206.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In: *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.*

Session 4.B.
Mondes professionnels

Du genre de discours aux pratiques langagières : usages de la question reformulée dans un corpus d'interactions en réunion de travail

Anouchka Divoux
ATILF UMR7118, Université de Lorraine
anouchka.divoux@univ-lorraine.fr

1 Introduction

L'étude que nous proposons dans cette communication s'intéresse à une pratique langagière particulière : la question. Analysées sous l'angle de la sociolinguistique des interactions verbales (André, 2015), les pratiques langagières sont appréhendées selon « leurs spécificités linguistiques, interactionnelles, pragmatiques et contextuelles » (*ibid*., p.2). Dans cette perspective, nous envisageons les pratiques langagières au regard d'une situation de communication toujours particulière. Ces pratiques s'insèrent dans différents niveaux de description des interactions verbales : les activités langagières, le genre de discours et la situation de communication (André, 2014). Ces différents niveaux de description sont interreliés et conditionnent les pratiques langagières utilisées dans une situation donnée, leurs formes et leurs objectifs pragmatiques (Hymes et Gumperz, 1972). Nous chercherons à montrer les implications du genre de discours « réunions de travail » sur cette pratique particulière et les activités qu'elle permet de réaliser. En effet, la question permet d'accomplir différentes activités, au-delà de celle qui vise à obtenir des informations de la part de son interlocuteur, telles que s'assurer du partage des connaissances, vérifier la bonne compréhension d'une information, etc. Si l'analyse des questions posées lors des réunions de travail met en lumière une variété d'objectifs pragmatiques, elle permet aussi de montrer le rôle de ces pratiques dans la réalisation de l'activité de travail. En prêtant une attention particulière au genre de discours (réunion de travail), nous traiterons ainsi des implications du travail dans l'interaction et de la forte intrication entre pratiques langagières et pratiques professionnelles (Boutet, 1995 ; Mourlhon-Dallies, 2007). Nous montrerons que les questions posées, tout en portant les traces de leur contexte de production, sont une des ressources langagières mobilisées par les participants pour s'assurer du bon déroulement de leurs activités.

2 Les questions en question

Dans le cadre de notre travail de thèse portant sur les pratiques permettant la construction de l'intercompréhension dans les réunions de travail, nous nous sommes intéressée à la question dans le sens où cette pratique permet d'accéder à l'information, de la partager ou encore de la préciser. Nous adoptons une définition à la fois pragmatique et interactionnelle de la question en considérant comme étant une question « tout énoncé qui se présente comme ayant pour finalité principale d'obtenir de son destinataire un apport d'information. » (Kerbrat-Orecchioni, 2008, p.86). D'une part, cette définition nous permet d'évacuer des formes qui sont traditionnellement liées aux questions, mais qui ne visent pas un apport d'information, c'est notamment le cas des questions rhétoriques. D'autre part, cette définition comprend des formes très éloignées

des questions prototypiques (questions en *est-ce que*, questions intonatives, etc.) mais qui visent tout de même un apport d'information. Les deux exemples suivants illustrent nos choix :

Exemple 1 :

L1 alors **à quoi ça sert ça que ce que je vous demande ici** si on était en production propre ça veut dire que je vais et là je demande en heures je vais peut-être avoir moi seize heures en pose + je vais peut-être avoir neuf heures l'ouvrage que je sous-traite il avait neuf heures à vivre en atelier + [...]

Exemple 2 :

L6 donc moi j'ai trois semaines pour fabriquer
L7 à peu près bah sauf que ça arrive le treize ça fait plus que deux semaines
///
parce que on on va essayer de sortir le
L6 pas trois + treize ↘
L7 bah une deux et après tu es au mois d'avril hein ah oui l'aut-
la semaine du treize tu veux dire ↘
L6 ouais
L7 ouais

Dans le premier exemple, l'énoncé présente l'adverbe interrogatif « quoi », prototypique d'une question. Pourtant, le locuteur ne laisse aucun moment opportun de prise de parole permettant à son interlocuteur de produire une réponse. Nous considérons donc que cet énoncé n'est pas une question[1]. Dans le second exemple, aucune marque explicite de la question n'est présente (adjectif ou adverbe interrogatif,

1. Cet énoncé est en fait une question introductive, soit un moyen argumentatif de structurer son discours tout en captant l'attention de ses interlocuteurs (Coveney, 1996).

Intérêt et limites des corpus oraux dans les formations linguistiques : le cas des corpus de réunions de travail en français pour la formation d'élèves ingénieurs chinois

Nian Liu
ICAR, Université Lumière Lyon 2
Nian.Liu@univ-lyon2.fr

1 Introduction

Se situant dans le cadre de la mobilité d'élèves ingénieurs chinois, la recherche présentée dans cette communication porte sur la maîtrise des interactions orales en réunions de travail en école d'ingénieurs. La Chine accorde en effet une attention particulière à la formation de ses ingénieurs et à son internationalisation. De nombreux étudiants chinois poursuivent ainsi des études dans des écoles d'ingénieurs françaises.

En France, les étudiants en école d'ingénieurs sont confrontés quotidiennement aux réunions de travail car les travaux de groupe et la collaboration constituent un des fondements de leur formation. Les élèves ingénieurs étrangers qui intègrent ces écoles ont donc besoin d'une préparation à ce type de situation de communication, mais il existe à ce jour très peu de matériels pédagogiques en FLE (français langue étrangère) traitant les interactions en réunion de travail.

2 Corpus et méthodologie

Notre recherche s'inscrit dans la démarche FOU (Français sur Objectif Universitaire), déclinaison de la démarche FOS (Français sur Objectif Spécifique) qui consiste à constituer des objets d'enseignement-apprentissage de la langue à partir des données collectées sur le terrain (Mangiante et Parpette 2004). Les travaux existants en FOU portent essentiellement, concernant l'oral, sur la réception des cours magistraux et la production des discours monologaux d'exposés ou soutenances (Mangiante et Parpette 2011, Carras et al. 2015), et n'abordent pratiquement pas les interactions de réunion de travail.

Dans notre travail, nous nous interrogeons sur la possibilité de créer des corpus oraux de réunions pour les utiliser, dans un programme de FOU, à destination des étudiants ingénieurs chinois. Nous avons pour cela suivi à l'École Centrale de Lyon un groupe d'élèves francophones qui réalisait un projet durant une année scolaire et avons filmé chaque séance hebdomadaire durant six semaines, soit environ 20 heures. L'objectif de cette constitution de corpus était d'en tirer des extraits pouvant servir de supports pédagogiques pour un programme de préparation linguistique en Chine.

3 Résultats

L'analyse du corpus ainsi constitué a révélé deux aspects majeurs, l'un méthodologique, l'autre discursif. Sur le plan méthodologique, le corpus s'est avéré indispensable pour informer

le concepteur de programme de FOU sur les contenus à traiter dans les séquences de formation linguistique. Sur le plan discursif, l'analyse a montré les limites de compréhensibilité des situations enregistrées. En effet, pour le spectateur extérieur que sont aussi bien le concepteur des séquences pédagogiques que les étudiants allophones à qui celles-ci sont destinées, les situations et les discours montrés par ces enregistrements comportent une importante opacité. Celle-ci est liée au poids important du contexte sur la compréhensibilité de ces interactions en réunions de travail. Les éléments d'opacité sont notamment une temporalité longue et difficilement maitrisable, le vécu partagé entre les interlocuteurs, et le caractère multimodal des interactions.

En ce qui concerne la temporalité, le temps des réunions de travail dépasse le corpus puisque les élèves se retrouvaient à travailler ensemble en dehors des séances de travail institutionnellement délimitées. La temporalité longue et la succession des réunions crée un vécu partagé important entre les interlocuteurs. Par exemple, dans une des vidéos collectées, un enseignant rappelle aux élèves qu'« il faut prendre rendez-vous le plus rapidement possible pour qu'on décide la date pour le RVP1 ». On ne peut pas comprendre cet énoncé sans connaître une information communiquée lors de la réunion précédente. En fait, tout au long de l'année, les élèves doivent organiser des rendez-vous de pilotage (RVP) pour rendre compte de l'avancement de leur projet à leurs tuteurs. Le premier rendez-vous de pilotage « RVP1 » doit avoir lieu avec les vacances de Noël. Par ailleurs, beaucoup d'interactions dans le corpus sont multimodales car les élèves travaillent souvent sur leurs ordinateurs et leurs échanges verbaux sont donc souvent liés aux contenus affichés sur les écrans, éléments de situation que nous n'avons pas pu capter.

Ces éléments qui échappent à la collecte de données créent l'opacité du corpus et rendent problématique la compréhension des données enregistrées a fortiori la création de supports pédagogiques à partir de ces données. Ce corpus dont nous avions au départ fait l'hypothèse qu'il pourrait servir de modèle dans le cadre d'une formation linguistique s'est finalement révélé relativement inopérant pour atteindre cet objectif.

Cela nous a amené à collecter un autre type de données avec lesquelles combiner le premier corpus pour pouvoir construire les supports pédagogiques visés : nous avons sollicité des interviews auprès des élèves et les avons interrogés sur leur expérience du projet. En effet, interviewer les acteurs des situations filmées permet d'expliciter les aspects absents du corpus collecté *in situ*. Ces interviews constituent un second corpus qui a pour fonction de compléter et d'élucider le corpus de réunions de travail. Par exemple, un élève a raconté un vécu partagé par les participants de leur projet : « Au début de l'année, on n'était pas d'accord sur les actions que nous voulions que les robots fassent. Il y avait une personne qui pensait pouvoir faire quelque chose avec les robots qui utilisaient une caméra de couleurs tandis que les autres pensaient que ça ne serait pas possible. Finalement, on a laissé la personne utiliser des caméras de couleurs. » Ce discours sollicité, provoqué par l'interview, concentre fortement les contenus des discours *in situ* complexes.

La combinaison de ces deux corpus de discours oraux *en situation* et de discours oraux *sur les situations*, permet d'élaborer des séquences de préparation linguistique aux réunions de travail.

Lorsqu'il s'agit d'interactions polylogales et fortement multimodales, les corpus de discours en situation constituent en eux-mêmes des outils limités pour l'élaboration de séquences de formation linguistique. En revanche, les discours sur les situations, sollicités à travers les interviews, jouent un rôle beaucoup plus important du fait de leur autonomie sémantique.

À cette dimension pragmatico-linguistique, s'ajoute la dimension culturelle : la culture universitaire et le fonctionnement des cursus diffèrent d'un pays à l'autre. Les élèves ingénieurs chinois ont plutôt l'habitude de travailler individuellement dans leur université d'origine et sont peu sensibilisés au travail collaboratif (Wang et Xiong, 2012). Les interviews montrent que travailler en groupe est un savoir-faire en cours d'acquisition pour les élèves ingénieurs de première année, qu'ils soient français ou étrangers. Il est important pour les élèves chinois de comprendre l'enjeu pédagogique et, ultérieurement, professionnel du travail en équipe. Les interviews peuvent sensibiliser les étudiants à ce type de situation.

Références bibliographiques

Carras, C. (2015). Les stratégies de collecte des données : Aspects institutionnels et déontologiques. *CCI Paris Ile-de-France*, (2), 20-36.

Carras, C., Gewirtz, O., & Tolas, J. (2014). *Réussir ses études d'ingénieur en français*. Grenoble : PUG.

Dufour, S., & Parpette, C. (2018). Le français sur objectif spécifique : La notion d'authentique revisitée. ILCEA. *Revue de l'Institut des langues et cultures d'Europe, Amérique, Afrique, Asie et Australie,* (32). `https://doi.org/10.4000/ilcea.4814`

Jouin-Chardon, E., Mondada, L., Niccolai, G. P., & Traverso, V. (2010). Contraintes technologiques sur les enregistrements de corpus et analyse des cadres de participation. *Pratiques*, (147-148), 53-81. `https://doi.org/10.4000/pratiques.1606`

Mangiante, J.-M., & Parpette, C. (2004). *Le Français sur Objectif Spécifique : De l'analyse des besoins à l'élaboration d'un cours* (1^{re} éd.). Paris : Hachette.

Mangiante, J.-M., & Parpette, C. (2011). *Le français sur objectif universitaire*. Grenoble : PUG.

Markaki, V. (2010). Filmer les réunions de travail en pratique : Réflexions sur l'enregistrement vidéo de phénomènes interactionnels complexes. *Presses universitaires de la Méditerranée*, (54-55), 283-298.

Ravazzolo, E., Traverso, V., Jouin-Chardon, E., & Vigner, G. (2015). *Interactions, dialogues, conversations : L'oral en français langue étrangère.* Vanves : Hachette, français langue étrangère.

Traverso, V. (2016). *Décrire le français parlé en interaction.* Paris : Éditions Ophrys.

Wang, L., & Xiong, Z. (2012). 工程师教育中团队精神的培养 < Le développement de l'esprit d'équipe dans la formation d'ingénieurs >. 北京航空航天大学 < *Beijing University of Aeronautics and Astronautics*>, *25*(4).

Session 5.A.
Corpus et enseignement

L'utilisation de corpus pédagogiques pour l'enseignement et la recherche : la question de l'acquisition lexicale

Heather Hilton [1], Ronald Peereman [2] et Michael Gauthier [1]
CRTT, Université Lyon 2
LPNC, Université Grenoble Alpes
heather.hilton@univ-lyon2.fr, ronald.peereman@univ-grenoble-alpes.fr, michael.gauthier.uni@gmail.com

1 Introduction

Dans les premières décennies du XX^{e} siècle, le psychologue américain Edward Thorndike a rassemblé le premier grand corpus constitué pour répondre à des besoins éducatifs : des textes de toutes sortes (littéraires, commerciaux, journalistiques, scolaires), contenant 4,5 millions de mots (Thorndike 1921). Il est difficile pour nous, en 2019, d'imaginer la minutie nécessaire pour élaborer manuellement à partir de ce grand corpus les premières listes de fréquence en anglais, rassemblées dans la célèbre série des *Teacher's Word Books* (Thorndike 1921 ; Thorndike 1931 ; Thorndike & Lorge 1944). Ces listes ont permis l'élaboration d'un syllabus lexical, pour l'apprentissage structuré et progressif de la lecture dans les écoles élémentaires américaines. Considérées comme « l'une des ressources scientifiques les plus utiles jamais développée » (Goodenough 1950 : 296), les listes de Thorndike ont également servi au développement d'outils psychométriques pour mesurer les connaissances lexicales et des niveaux de compétence en lecture.

En collaboration avec Thorndike (Fawcett, Palmer, Thorndike & West 1936), Michael West a élargi et ajusté le corpus de base (le portant à 5 millions de mots), pour identifier les mots les plus utiles (« *of greatest general service* ») pour l'apprentissage de l'anglais langue étrangère ou seconde par des apprenants adultes (West 1936 ; West 1953). A la suite de ce travail didactique, des initiatives parallèles ont été entreprises dans d'autres pays européens après la Seconde Guerre Mondiale, dans un contexte de mobilité internationale et de promotion soutenue de l'enseignement des langues (par exemple, Gougenheim et al. 1956, pour le français langue étrangère). Ces travaux lexicologiques ont fourni aux auteurs des manuels de langues un syllabus lexical, permettant l'introduction progressive des mots selon la fréquence de leur utilisation, et donc une graduation des supports (textes et enregistrements), pour un apprentissage structuré et donc optimisé.

Lors de la « révolution » communicative des années 1980 (en Europe et aux Etats-Unis), la notion d'un programme lexical structuré fut abandonnée en didactique des langues (voir, par exemple, Auroux 1985, pour une vision résolument anti-lexicale de la compétence communicative). L'attention méthodologique depuis cette époque est focalisée sur les « activités langagières » en classe de langue (compréhension, expression, interaction), et le programme est basé sur des « actes de parole » ou différentes fonctions interactionnelles du langage. Cette centration didactique sur l'utilisation du langage a eu comme effet de déstructurer la programmation des éléments formels de la langue à apprendre (vocabulaire, morphologie, syntaxe, prononciation) :

manuels et enseignants abordent les formes en fonction des besoins communicatifs de telle situation interactionnelle, et non plus selon des critères linguistiques (comme la fréquence d'un mot ou d'une forme grammaticale : Meara 1980 ; Swan 1985 ; Arnaud et al. 1985 ; Nordlund 2016 : 48-50). La dernière décennie en France a vu le retour des préoccupations lexicales dans les classes de français langue maternelle (et notamment l'importance du vocabulaire dans l'apprentissage de la lecture, Dehaene 2011), mais ce renouveau d'intérêt pour le vocabulaire n'est pas encore reflété dans les textes qui régissent l'enseignement des langues vivantes en France[1].

2 Corpus et méthodologie

2.1 Corpus

Dans ce contexte, un groupe de chercheurs dans quatre universités françaises a élaboré un projet visant à analyser de plus près l'acquisition lexicale dans les classes de langue vivante en France, à l'école élémentaire et au collège. La première tâche de ce projet porte sur l'analyse d'un grand corpus de manuels utilisés pour enseigner l'anglais dans les quatre années du collège (60 manuels, 15 pour chaque année de collège), à l'image de la base lexicale *Manulex*, compilée sur 54 manuels scolaires utilisés dans l'enseignement élémentaire en France (Lété *et al.* 2004). Trente-deux manuels ont été numérisés jusqu'ici, générant un corpus de 725 720 mots, avec en moyenne environ 181 000 mots par année de collège. Un deuxième volet de cette tâche consiste à filmer quatorze leçons d'anglais en collège (trois ou quatre leçons par année de collège), générant un petit corpus oral parallèle, qui complète le corpus écrit et permet quelques comparaisons lexicologiques entre manuels et leçons.

2.2 Méthodologie

Dans cette communication, nous présenterons la méthodologie utilisée pour numériser les manuels et établir des listes de mots selon leur fréquence et leur fonction grammaticale : les listes sont annotées et lemmatisées selon CLAWS (Garside & Smith 1997) et *Stanford NLP* (Toutanova *et al.* 2003). Nous résumerons également les méthodes et outils utilisés pour transcrire le corpus des leçons avec le logiciel EXMARaLDA (Schmidt *et al.* 2014) et l'étiqueter avec *WMatrix* (Rayson 2008).

3 Résultats

Nos premières analyses du corpus écrit révèlent une grande et surprenante disparité lexicale entre les manuels. Alors que 20 000 types sont dénombrés (incluant 20% de sigles et noms propres –un lexique très étendu, pour les niveaux européens A2 et B1 visés en collège), seulement 3500 types (17,5%) sont partagés entre les quatre niveaux scolaires. Ce ratio n'est que légèrement plus élevé (24,27%) lorsque l'analyse porte sur les lemmes, sans prise en compte des noms propres et sigles. De façon inattendue, les manuels d'une même année (les huit manuels

1. Les programmes du Socle commun insistent sur l'importance du vocabulaire 80 fois dans les sections consacrées à l'enseignement du français, de l'histoire, des arts plastiques... mais une seule fois dans les sections dédiées à l'enseignement des langues vivantes (Ministère de l'éducation nationale 2015).

de 5^{e} analysés jusqu'ici, par exemple) ne partagent que 5,5% de leurs types ; dans un contexte pédagogique structuré, on s'attendrait à un recouvrement nettement plus élevé (Nordlund 2016). Plus surprenant encore, 186 mots (7%) du *New General Service* List (Browne *et al.* 2013) sont absents de ces manuels de collège (des mots comme *ally*, *overall*, *sufficient*, *income*, *strengthen*, *tire*...) ; pourtant, la totalité des 2801 mots de cette liste de base (les mots incontournables pour une utilisation réceptive ou productive de l'anglais) devrait logiquement figurer –avec une fréquence élevée –dans des supports à ces niveaux élémentaires (Nordlund 2016, pour des résultats semblables dans des manuels d'anglais utilisés en Suède). De façon plus rassurante, la grande majorité (85 %) des mots entendus en classe se retrouve dans les manuels du même niveau, mais nous trouvons aussi (à l'image du corpus écrit) que ce taux de recouvrement lexical est quasi identique, quel que soit le niveau des manuels que l'on compare aux leçons.

Les premières analyses de ce corpus de manuels d'anglais donnent donc des résultats inattendus, qui soulèvent des questions didactiques de fond. En conclusion, nous évoquerons quelques retombées d'un programme lexical diffus et sans doute trop étendu : le faible niveau des élèves français en anglais langue étrangère (compréhension de l'oral et de l'écrit, expression écrite), selon l'étude européenne *Surveylang* (European Commission 2012 ; Beadle & Scott 2014) ; leurs connaissances lexicales limitées en anglais L2 à l'arrivée dans l'enseignement supérieur (Hilton 2019 : 25). Nous mentionnerons également les démarches expérimentales basées sur les listes du corpus, qui auront comme objectif de mesurer l'émergence des connaissances lexicales en anglais L2 chez des collégiens français (tâche de décision lexicale). Notre communication illustrera donc deux utilisations possibles des corpus : didactique (la programmation des contenus lexicaux en langue étrangère) et expérimentale (la conception d'outils pouvant mesurer les acquis lexicaux).

Références bibliographiques

Arnaud, P. J. L., Béjoint, H. & Thoiron, P. (1985). A quoi sert le programme lexical ? *Les Langues Modernes 3/4*, 72-85.

Auroux, S. (1985). Le droit à l'oubli : réponse à Arnaud, Béjoint et Thoiron. *Les Langues Modernes 3/4*, 86-91.

Beadle S, and Scott D (2014). *Languages in Education and Training : Final Country Comparative Analysis*, Report n° J9241. European Commission.

Browne, C., Culligan, B. & Phillips, J. (2013). *New General Service List*. Tokyo.

Dehaene, S. (dir ; 2011). *Apprendre à lire : Des sciences cognitives à la salle de classe*. Paris : Odile Jacob.

European Commission (2012) *First European Survey on Language Competences : Final Report.* Brussels : European Commission Education and Training Division.

Faucett, L. W., Palmer, Thorndike, E. L., & West, M. P. (1936). *Interim Report on Vocabulary Selection for the Teaching of English as a Foreign Language*. London.

Garside, R. & Smith, N. (1997). A hybrid grammatical tagger : CLAWS4. Dans R. Garside, G. Leech & A. McEnery (dir), *Corpus annotation : Linguistic information from computer text corpora.* London : Longman, 102-121.

Goodenough, F. L. (1950). Edward Lee Thorndike : 1874-1949. *The American Journal of Psychology* 63(2), 291-301.

Gougenheim, G., Michea, R., Rivenc, P. & Sauvageot, A. (1956). *L'élaboration du français élémentaire : Étude sur l'établissement d'un vocabulaire et d'une grammaire de base.* Paris : Didier.

Hilton, H. E. (2019). *Sciences cognitives et didactique des langues, Rapport d'expertise.* Paris : CNESCO.

Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). MANULEX : A grade-level lexical database from French elementary-school readers . *Behavior Research Methods, Instruments, & Computers, 36*, 156-166.

Meara, P. M. (1980). Vocabulary acquisition : A neglected aspect of language learning. *Language Teaching and Linguistics Abstracts, 13*, 221-46.

Ministère de l'éducation nationale (2015). *Programmes pour le cycle 2, 3, 4.* Paris : MEN.

Nordlund, M. (2016). EFL textbooks for young learners : a comparative analysis of vocabulary. Education Inquiry•, 7(1), 47-68.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics.* 13(4), 519-549.

Schmidt, T., & Wörner, K. (2014). EXMARaLDA. In U. Gut, J. Durand & G. Kristofferse (dir), *Handbook on Corpus Phonology.* Oxford : Oxford University Press, 402-419.

Swan, M. (1985). A critical look at the Communicative Approach. *ELT Journal, 39*(1-2), 2-12 ; 76-87.

Thorndike, E. L. (1921, 1931). *A Teacher's Word Book : the Twenty Thousand Words Found Most Frequently in General Reading for Children and Young People.* New York : Teachers College.

Thorndike, E. L., & Lorge, I. (1944). *The Teacher's Word Book of 30,000 Words*. New York : Teachers College.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, 252-259.

West, M. P. (1936, 1953). *A General Service List of English Words.* London : Longman.

Le point de vue de l'apprenant dans une approche sur corpus en cours de rédaction scientifique en anglais : typologie des besoins, questions et actions, correspondance questions/actions/outils.

Sylvain Perraud
ATILF, Université de Lorraine
LIDILEM, Université Grenoble Alpes
sylvain.perraud@univ-grenoble-alpes.fr

1 Introduction

Un défi majeur du recours à une approche sur corpus, ou « par les données » (data-driven learning, DDL) dans l'apprentissage des langues, est le développement de connaissances sur la mise en œuvre des outils de consultation de corpus par les apprenants. Une telle caractérisation est essentielle non seulement en situation d'apprentissage guidé ou semi-guidé, mais aussi (voire surtout) une fois que les apprenants sont livrés à eux-mêmes.

Une meilleure prise en compte du point de vue des utilisateurs, visant à coller au mieux à leurs besoins réels et à leurs manières d'apprendre, a en effet plusieurs implications majeures en didactique des langues, et en particulier dans ses aspects suivants :

- rationalisation du choix des outils par les formateurs ;
- optimisation du développement d'outils ;
- adaptation des programmes et de l'organisation des formations ;
- affinement du développement de ressources pédagogiques complémentaires aux outils.

L'étude de l'utilisation des corpus par les apprenants a été précédemment caractérisée, pendant la formation (Gilmore 2009 ; Pérez-Paredes 2013 ; Yoon 2004) mais également en dehors (Chen 2018 ; Charles 2014), avec notamment pour objectif la conciliation de l'appoint lexico-grammatical (mise en cohérence de la production avec la langue telle qu'usuelle) et du travail d'exploration phraséologique permettant de s'approprier une gamme de routines discursives correspondant aux fonctions rhétoriques récurrentes dans le genre cible (Charles 2007).

Peu d'études proposent cependant un suivi dans le temps, pendant et après la formation, de l'utilisation d'une gamme d'outils par les apprenants, combinant actions menées et retours sur les outils utilisés.

Nous nous intéresserons ici, en particulier, aux questions suivantes :

- quels critères guident le choix d'outils par les apprenants ?
- comment s'en servent-ils ?

- qu'y cherchent-ils ?
- le cas échéant, les informations recueillies sont-elles exploitables, complètes ?
- quelles questions éventuelles se prêtent peu (ou pas) à une interrogation de corpus ?
- quels commentaires et critiques formulent-ils sur les outils ?

2 Corpus et méthodologie

La mise en place d'une approche sur corpus dans une formation de rédaction d'articles scientifiques pour doctorants du secteur Sciences-Technologie-Santé a été l'occasion d'une telle caractérisation systématique de l'utilisation par les apprenants d'un éventail de cinq outils présentés en cours : Antconc (Anthony 2017), Coca (Davies 2008), Hyper Collocation (Maruta 2018), Netspeak (Potthast 2010), et ScienQuest (Falaise 2011).

Il est proposé de dresser ici un bilan du déploiement initial de cette approche sur une cohorte de quatre groupes de 15, soit 60 étudiants.

Le but de l'étude présentée n'est pas de caractériser l'efficacité du dispositif. Il a été montré par ailleurs que le recours à une approche type DDL a un effet net positif dans la grande majorité des cas (Boulton 2017). C'est cette efficacité, en particulier dans l'apprentissage de l'écrit, qui a guidé le choix initial d'une approche semi-inductive sur corpus dans la refonte de la formation considérée.

L'objectif central de cette étude est bien, en revanche, de caractériser le travail autonome que font les étudiants, suivant trois axes principaux : outils, besoins et actions.

3 Résultats

Plus d'un millier de recherches (nombre total 1153) effectuées par les étudiants (« requêtes ») ont été recensées dans la première cohorte de 60 étudiants concernés en 2018. Ces requêtes ont utilisé principalement, mais non-exclusivement, les outils présentés en cours. Leur typologie fait clairement apparaître un nombre restreint de catégories. Hormis les recherches de synonymes, qui sont en dehors du champ direct de l'approche par corpus (235 requêtes, soit 20,4% du total), la quasi-totalité des requêtes effectuées par les apprenants peuvent être regroupées suivant cinq types principaux : (1) colligations (401 requêtes ; 34,8%) ; (2) collocations (208 ; 18,0%) ; (3) contextualisation (vérification de l'utilisation d'un terme ou d'une expression en contexte) [201 ; 17,5%]; (4) usualité (vérification de la conformité d'un élément linguistique avec la réalité de l'usage) [78 ; 6,8%]; (5) comparaison (de deux ou plusieurs options afin d'effectuer un choix) [29 ; 2,5%].

Les analyses effectuées révèlent des tendances de fond, en particulier la prépondérance de requêtes lexico- grammaticales du domaine scientifique transdisciplinaire ou encore la distribution très large de certains types de questions malgré la variété d'horizons thématiques des étudiants concernés.

La plupart des tendances observées dans les types de requêtes et le choix des outils montrent une dépendance temporelle marquée. Nous tenterons de mettre en perspective les évolutions considérées en nous référant notamment sur des travaux aux objectifs similaires (Charles 2014 ; Chen 2018) et proposerons des interprétations possibles, en considérant plus largement les différents paramètres corrélés à l'aspect temporel, entre autres : échéances, appropriation progressive des outils, prise de confiance, affinement des besoins avec l'avancée du processus rédactionnel, et influence des dynamiques de travail en classe.

Parmi les facteurs-clés rentrant en ligne de compte dans les tendances observées, nous examinerons en particulier niveau en langue, le contexte de consultation des outils, l'expérience rédactionnelle, l'année de thèse, ou encore la langue maternelle des participants.

L'influence d'éléments plus complexes, introduisant des biais potentiels, sera également discutée, en particulier certaines caractéristiques des outils utilisés et le contexte de leur prise en main initiale, mais aussi différents types de contraintes extérieures, notamment temporelles (pressions liées à des échéances ou au déroulement du cours) et socio-psychologiques (sous-déclaration potentielle de certain types de requêtes, pouvant provenir notamment d'une gêne vis-à-vis de la perception du formateur ou des pairs).

Divers facteurs se conjuguent également pour complexifier l'exploitation et l'interprétation des données, notamment la diversité de profil du public : niveau de langue, langue maternelle, année de thèse, expérience et habitudes rédactionnelles, échéances éventuelles de publication ;la diversité de motivations de suivi de la formation ; des variations d'assiduité aux sessions présentielles entre participants ; des disparités d'adhésion aux activités proposées, notamment dans le travail autonome ; la diversité des contextes d'utilisation des outils, (e.g. décalage temporel entre rédaction et requête).

Ce nombre élevé de sources potentielles de biais plaide pour un complément de l'étude globale par quatre études de cas permettant d'apporter un éclairage supplémentaire dans l'analyse d'un certain nombre de situations-types.

Parmi les perspectives proposées, sont présentées les principales modifications implémentées, sur la base des enseignements tirés des données 2018, dans une deuxième série de ces formations, ayant eu lieu en 2019 et ayant donné lieu nouvelle campagne de collecte de données.

Références bibliographiques

Boulton, A. (2010). Learning outcomes from corpus consultation. In F. Serrano Valverde, M. Moreno Jaén, & M. Calzada Pérez (Eds.), *Exploring new paths in language pedagogy : Lexis and corpus-based language teaching* (pp. 129–144). London : Equinox.

Boulton, A. & Cobb, T. (2017) Corpus Use in Language Learning : A Meta-Analysis. *Language Learning*. `https://onlinelibrary.wiley.com/doi/abs/10.1111/lang.12224`

Charles, M. (2007). Reconciling top-down and bottom-up approaches to graduate writing : Using a corpus to teach rhetorical functions. *Journal of English for Academic Purposes*, 6(4), 289–302.

Charles, M. (2014). Getting the corpus habit : EAP students' long-term use of personal corpora. *English for Specific Purposes*, 35. 30–40. 10.1016/j.esp.2013.11.004.

Chen, M. & Flowerdew, J. (2018). Introducing data-driven learning to PhD students for research writing purposes : A territory-wide project in Hong Kong. *English for Specific Purposes*, 50. 97-112

Flowerdew, L. (2010). Using corpora for writing instruction. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 444–457). London : Routledge.

Gilmore, A. (2009). Using online corpora to develop students' writing skills. *ELT Journal*, 63(4), 363–372.

Pérez-Paredes, P., Sánchez-Tornel, M. Alcaraz Calero, J. (2013). Learners' search patterns during corpus-based focus-on-form activities. *International Journal of Corpus Linguistics*, 17(4), 482–515

Yoon, H. (2008). More than a linguistic reference : The influence of corpus technology on L2 academic writing. *Language Learning and Technology*, 12(2), 31–48.

Yoon, H. & Hirvela, A. (2004). ESL student attitudes towards corpus use in L2 writing. *Journal of Second Language Writing*, 13(4), 257–283.

Références logicielles

"ANTCONC" : Anthony, L. (2017). AntConc (Version 3.5.0) Tokyo, Japan : Waseda University. <laurenceanthony.net/software>

"COCA" : Davies, Mark. (2008) The Corpus of Contemporary American English (COCA) : 560 million words, 1990-present. <corpus.byu.edu/coca>

"HYPER COLLOCATION" : Maruta, I. (2018). <hypcol.marutank.net> "NETSPEAK" : Potthast, M., Trenkmann, M. & Stein, B. (2010). Netspeak - Assisting Writers in Choosing Words. *Advances in Information Retrieval : 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK* (Proceedings). Bauhaus-Universität, Weimar, Germany. <netspeak.org>

"SCIENQUEST" : Falaise, A., Tutin, A. & Kraif, O. (2011) Définition et conception d'une interface pour l'exploitation de corpus arborés pour non-informaticiens : la plateforme ScienQuest du projet Scientext. TAL, 52(3), 103-128. <corpora.aiakide.net/scientext19>

Des corpus d'interactions à l'enseignement du français parlé : objectifs et ressources de la plateforme CLAPI-FLE

Biagio Ursi [1,2] et Carole Etienne [2]
[1]ATILF, CNRS / Université de Lorraine
[2]ICAR, CNRS / ENS de Lyon / Université Lyon 2
ursi1@univ-lorraine.fr, carole.etienne@ens-lyon.fr

1 Introduction

L'équipe LIS du laboratoire ICAR collecte, documente et transcrit depuis maintenant une quarantaine d'années des corpus oraux d'interactions écologiques recueillies *in situ*, sans aucune intervention du chercheur dans leur déroulement, pour mener à bien ses analyses et ses travaux de recherche en interaction (Traverso, 2016). Elle s'est intéressée dès les années 2000 à la mise à disposition de ses données à la communauté scientifique en linguistique, et plus généralement dans d'autres disciplines, en proposant la plateforme en ligne CLAPI[1] (Groupe ICOR, 2016), qui rend accessible les corpus mais également un ensemble d'outils permettant de les exploiter.

2 La plateforme CLAPI

Les corpus oraux de CLAPI concernent des interactions en milieu professionnel et privé, dans des contextes variés comme des achats dans des commerces, des réunions professionnelles, des apéritifs ou des repas, des préparations de repas, des invitations téléphoniques, des prises de rendez-vous, des conversations en ligne, des sessions de jeux, des consultations médicales, des visites guidées. Les situations documentées s'enrichissent au fur et à mesure des travaux de l'équipe et des nouveaux terrains de recherche, alimentant ainsi régulièrement la base de données CLAPI, qui héberge à ce jour une centaine d'enregistrements d'une durée totale de 63h.

3 La plateforme CLAPI-FLE : des corpus d'interactions à l'enseignement du français parlé

Dans une perspective similaire de partage et d'adaptation de ses corpus, un réseau de partenaires didacticiens et enseignants a été constitué pour étudier ensemble si certains des corpus de la base CLAPI ou des travaux de recherche de l'équipe LIS-ICAR pourraient représenter des ressources pour l'enseignement du français ou de la linguistique française, établissant ainsi une passerelle entre la didactique des langues et l'analyse des interactions (Ravazzolo et al., 2015). À ce jour, ce réseau regroupe des collègues en France (ENS de Lyon, Lyon 2, Paris 3, Paris 5, Aix-Marseille, Rennes) et à l'étranger (Italie, Espagne, Suisse, Suède, Chine, Inde, Vietnam). Cette démarche s'appuie sur le fait que les corpus oraux de CLAPI sont des situations ordinaires, semblables à celles qu'un apprenant de français rencontrerait au quotidien dès son arrivée dans un pays francophone. L'utilisation de ces corpus permet ainsi de se rapprocher de la situation d'immersion vivement recommandée dans l'apprentissage d'une nouvelle langue (Lhote, 1995).

1. http://clapi.icar.cnrs.fr

En parallèle, nous avons consulté nos collègues engagés dans la même démarche de constitution de ressources à partir de corpus oraux, travaillant sur le projet IPFC - *Interphonologie du Français Contemporain*, coordonné par Isabelle Racine (Université de Genève, Suisse) et Sylvain Detey (Université de Waseda, Japon) ou sur le projet FLEURON - *Français Langue Étrangère Universitaire : Ressources et Outils Numériques*, basé sur des enregistrements utiles aux étudiants étrangers en France dans le cadre du Français sur Objectifs Universitaires (FOU) et développé par Virginie André du groupe LTF Langage, Travail et Formation au laboratoire ATILF (Université de Lorraine). Cette collaboration s'est poursuivie par des journées d'études en 2015 et en 2017, ainsi que des contributions à des colloques et tables rondes (Zay et al., 2016 et André & Etienne 2018, entre autres).

Dans notre communication, nous présenterons cette approche méthodologique qui nous a permis de répondre aux attentes des enseignants, en concertation avec notre réseau de chercheurs en didactique et en linguistique de l'oral, en concevant des ressources innovantes à partir de nos corpus oraux et des résultats de nos travaux en interaction. En effet, si le lien entre la linguistique de corpus et la didactique des langues a été établi et encouragé par de nombreux travaux et initiatives (Debaisieux, 2009 ; Boulton & Tyne, 2014 ; André, 2018), il n'en demeure pas moins que les corpus oraux ont bien du mal à franchir la salle de classe, les enseignants leur préférant les émissions de radio, les séries télévisées ou bien les données orales construites des méthodes de langue. Ces dernières ressources n'offrent bien souvent qu'une seule représentation parmi la variété des réalisations possibles à l'oral, lui donnant ainsi une valeur normative, alors qu'elle est en fait très éloignée des productions réellement attestées (Giroud & Surcouf, 2016). Selon l'approche interactionniste, la langue orale est conçue comme un ensemble de ressources dynamiques et en perpétuelle adaptation, leurs usages relèvent de la compétence interactionnelle. Cette compétence est « *située, contextualisée* dans la mesure où elle est structurée (tant dans son développement que dans sa mobilisation) en réponse à l'accomplissement local des activités, leur articulation aux activités d'autrui et la mobilisation de méthodes –façons systématiques de faire, acquises à travers les processus de socialisation » (Pekarek Doehler 2006 : 39). Cette compétence est aussi collective dans la mesure où elle s'appuie sur les expériences partagées avec d'autres acteurs sociaux, elle est co-construite à travers la variété des pratiques langagières et des usages situés qui sont expérimentés à travers les échanges avec des locuteurs divers. La difficulté de l'apprenant vient justement de cette variabilité à laquelle il se trouve confronté sans véritable préparation.

Pour vérifier si nos corpus pouvaient être compris par des apprenants, un premier projet exploratoire a été réalisé en collaboration avec Anita Thomas et Jonas Granfeldt, didacticiens du Département de français à l'Université de Lund (Suède), un processus expérimental a été construit en soumettant quatre extraits de différents niveaux de difficulté à des apprenants, en leur demandant d'expliquer ce qui se passait et d'identifier les séquences de désaccord. Les résultats obtenus ont été au-delà des préconisations du CECR, qui établit les compétences attendues en matière de compétence d'interaction par niveau de langue, et il s'est avéré que la naturalité des données facilitait la compréhension des séquences par la prise en compte du contexte, voire des bruits ambiants (Thomas et al., 2016).

À partir de ce résultat et dans le cadre d'une collaboration avec Florence Mourlhon-Dallies, Martina Ronci et Sabine Henri (Université Paris 5), l'équipe LIS a initié la conception de la plateforme CLAPI-FLE[2], en proposant une quarantaine d'extraits décrits, transcrits et didactisés qui permettent différentes exploitations suivant les objectifs pédagogiques et les niveaux des apprenants (Alberdi et al., 2018). Les divers contextes représentés permettent d'appréhender les différentes composantes d'une interaction orale pour comprendre son organisation et les procédés que les locuteurs mettent en œuvre afin d'accomplir des objectifs langagiers spécifiques (Boulton & Tyne, 2014). La prise en compte de ces procédés va au-delà de la simple connaissance du lexique et de la grammaire, à laquelle trop de méthodes de langues se limitent encore.

En parallèle de ces extraits, des collections ont été proposées pour illustrer par des exemples certaines spécificités de l'oral (dislocations, imparfait de politesse, futur simple) ou certaines tournures (”c'est vrai”, ”le truc c'est que”). Elles seront complétées par les illustrations de certaines actions langagières comme expliquer, refuser, inviter ou évaluer.

Au-delà de la mise à disposition d'extraits de données, nous avons souhaité transposer les résultats de certains de nos travaux de recherche en ressources pour l'enseignement. Nous avons ainsi expliqué de quelle manière certaines fonctions langagières étaient réalisées et quels procédés étaient utilisés par les locuteurs sur le plan du lexique, des marqueurs de l'oral, de la grammaire, du registre, de la prosodie, de la multimodalité ou de l'approche multiculturelle. En concertation avec notre réseau de partenaires et dans le cadre d'une collaboration avec Élodie Oursel et Carolina León Roa (Université Paris 8), nous avons conçu des ”fiches explicatives” sur les remerciements, le discours rapporté, les atténuateurs, les questions et certaines expressions complexes (*trop, quand même*). Ces fiches proposent une typologie des attestations, des explications de chacune d'entre elles et une rubrique ”Pour aller plus loin”, incluant des usages moins fréquents et une bibliographie.

En présentant ces ressources et ces extraits, les enseignants nous ont demandé à plusieurs reprises si nous pouvions les aider à enseigner cette compétence d'interaction avec laquelle ils ne se sentaient pas toujours à l'aise. Pour répondre à cette attente, nous avons complété la plateforme en ajoutant un volet ”analyse interactionnelle des extraits” pour les expliciter et les rendre plus accessibles du point de vue de l'organisation des séquences et de l'identification des actions langagières.

Notre communication retracera les principales étapes de notre approche méthodologique et illustrera l'articulation de ces différentes ressources pour différents objectifs d'enseignement et différents publics.

Références bibliographiques

Alberdi, C., Etienne, C. & Jouin-Chardon, E. (2018). Les apports des corpus d'interactions naturelles en situation de classe : enjeux et pratiques. *Action didactique*, 1, 55-70.

2. `http://clapi.icar.cnrs.fr/FLE`

André, V. (2018). Nouvelles actions didactiques : faire de la sociolinguistique de corpus pour enseigner et apprendre à interagir en français langue étrangère. *Action didactique*, 1, 71-88.

André, V. & Etienne, C. (2018). Apprendre le français parlé en interaction avec les ressources Fleuron et Clapi-FLE. *Journée Ressources linguistiques et didactique des langues*, Rennes, <halshs-01958673>.

Debaisieux, J-M. (2009). Des documents authentiques oraux aux corpus : un défi pour la didactique du FLE. *Mélanges CRAPEL,* 31, 35-56.

Boulton, A. & Tyne, H. (2014). *Des documents authentiques aux corpus : démarches pour l'apprentissage des langues.* Paris : Didier.

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, *SHS Web of Conferences*, 27, `https://doi.org/10.1051/shsconf/20162707017`.

Groupe ICOR (Baldauf-Quilliatre, H, Colon de Carvajal, I., Etienne, C., Jouin-Chardon, E., Teston-Bonnard, S., Traverso, V.) (2016). CLAPI, une base de données multimodale pour la parole en interaction : apports et dilemmes. In Avanzi M., Béguelin M.-J. & Diémoz F. (dir.), Corpus 15 –numéro thématique sur *Corpus de français parlés et français parlés des corpus*, `https://journals.openedition.org/corpus/2991`.

Lhote, E. (1995). *Enseigner l'oral en interaction*. Paris : Hachette.

Pekarek Doehler, S. (2006). Compétences et langage en action. *Bulletin suisse de linguistique appliquée*, 84, 9-45.

Ravazzolo, E., Jouin, E., Traverso, V., Vigner, G. (2015). *Interactions, dialogues, conversations : l'oral en français langue étrangère.* Paris : Hachette.

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.

Traverso, V. (2016). *Décrire le français parlé en interaction*. Paris : Ophrys.

Zay, F., André, V., Cortier, C., Etienne, C., Pêcheur, J. (2016). Authenticité et didactisation : les documents et les situations à l'épreuve de la salle de classe –Table ronde, *Colloque Variation, plurilinguisme et évaluation en FLE*, Genève, <halshs-01357145>.

Session 5.B.
Pragmatique et énonciation

Factuality in texts: A new project, a new tool, a new corpus

Glòria Vázquez [1], Hortènsia Curell [2], Ana Fernández-Montraveta [2], Leyre Barrios [1] et Irene Castellón [3].

[1]Universitat de Lleida

[2]Universitat Autònoma de Barcelona

[3]Universitat de Barcelona

gvazquez@dal.udl.cat, hortensia.curell@uab.cat, ana.fernandez@uab.cat, leyre.barrios@udl.cat, icastellon@ub.edu

1 Introduction

The categorization of events with respect to their factual status is an area of great interest in the field of Corpus Linguistics and Natural Language Processing (NLP). The most habitual approach to annotate factuality in a corpus is to consider that the degree of certainty of an event is related to the way in which the event is presented by the writer.

The ultimate objective of the project that we present (TAGFACT), which is a year and a half into its development, is to create a system for the automatic annotation of the factual values of events expressed in Spanish texts, using linguistic knowledge. In Spanish, this issue has not been dealt with in much depth, and what little has been done has been developed based on statistical processes. In our case, we will use linguistic knowledge and we will focus on journalistic texts.

The first step was to create a corpus that we are annotating manually with a twofold objective. First, we will use it to study factuality marks in Spanish, and, hence, infer the knowledge that can be formalized to carry out the automatization later. Second, we will turn it into a Gold Standard that we will use as a test base in the second phase.

With that purpose in mind, we carried out the linguistic analysis of a small sample of Spanish texts extracted from the corpus created specifically for the project. On the basis of existing tagsets from other projects and our own analysis, we elaborated a proposal of the set of labels that will be used in the project to annotate factuality. Besides, we designed and implemented an annotation editor that is powerful enough to label widely different aspects during the process of annotation. In this paper, we describe the design of both resources (the corpus and the editor), and we will leave the argumentation in support of our proposal of labels for future work.

2 Factuality in Corpus Linguistics and NLP

One of the ground breakers in the annotation of factuality in texts is FactBank (Saurí & Pustejovsky, 2009). This factually annotated corpus was an innovative proposal for the representation of factual information in English. It contains 9488 events (208 files) manually annotated according to different sources. Within the framework of her PhD thesis, Saurí (2008) also implemented a factuality annotator (De Facto). Although this tool contains an algorithm to annotate factuality automatically, some of the knowledge modules used were created manually, so labelling cannot be considered to be an automatic process.

From the moment of the creation of FactBank and De Facto, the interest in factuality is established within the NLP community. Various authors have drawn on FactBank (2009) for the annotation of various corpora, although some have modified certain values. Thus, we have Diab et al. (2009), Soni et al. (2014), van Son et al. (2014) and Lee et al. (2015) for English; Matsuyoshi et al. (2010) and Narita et al. (2013) for Japanese; Minard, Speranza, & Caselli, (2016) for Italian; and Wonsever et al. (2016) for Spanish. As for the latter, these authors built one of the few corpora annotated with factual and temporal information in Spanish. In other work, opposite to the framework used in FactBank and De Facto, the authors have considered factuality as linked to the knowledge of the world (Marneffe et al., 2012). One of the fields in which more work has been carried out in relation to the annotation of factuality is biomedicine (Morante et al. (2010) and Velupillai (2011), among others).

3 TAGFACT corpus and editor

At present, we are compiling and manually annotating a part of the corpus, which will constitute the Gold Standard, that is, the part of the corpus manually corrected that will allow us to assess the TAGFACt corpus at the end of the automatic annotation process.

The Gold Standard TAGFACT corpus contains 74 pieces of news from newspapers published in peninsular Spanish, organized in 29 groups As a general rule, each group contains 3 different pieces dealing with the same incident narrated by newspapers with different ideology (*La Razón, El Periódico and El Diario*). However, when it was not possible to find one of the pieces in these media, it was taken from another one with similar characteristics (*ABC, Público, La Vanguardia or 20 Minutos*, among others). In each piece of news, the following information is codified:

- Metadata: name of the newspaper, section, date, author, news URL and geographical location
- Data: news title, subtitle and text
- Media: images and Twitter comments (only informative)

The most frequent sections are politics and international, although there are also news from society and sports. The total volume of words in the corpus is approximately 30000, grouped into 930 sentences, which contain 5819 predicates.

The task flux in the process of annotation of the sentences in the corpus includes the introduction of the news, and the automatic linguistic analysis of the sentence (Freeling - Padró et al. 2012), a process that allows us to segment the pieces of news into sentences and predicates, and, finally, the manual annotation.

As already mentioned, an annotating editor was created to carry out this flux. With this editor it is possible to create and manage several corpora by means of an interface that allows the user to create a new corpus, to codify data, to edit the text entered and to send it for analysis. That is, once the user has corrected the text, if needed, there is an option to send the sentences

to Freeling to be syntactically and semantically analyzed. The piece of news is received from Freeling segmented into sentences, and the information we retrieve for each sentence, which can be seen in the editor, is the list of the predicates and their corresponding arguments (in figure 1 we see the identification of 4 arguments for the predicate *llegó*).

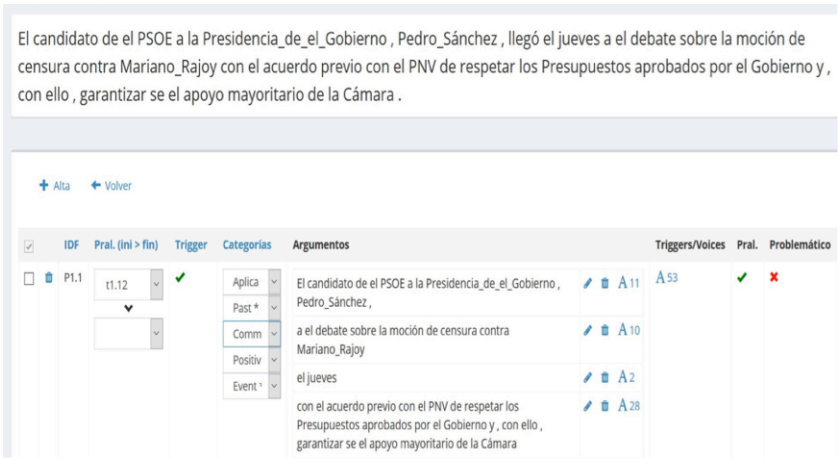


Figure 1: Editor interface

This structure received from Freeling is then corrected manually. First, the segmentation into sentences has to be revised; second, the predicates and arguments identified by Freeling (together with their scopes) must be validated. Then, the categorization of each predicate regarding its factual status is carried out (in figure 1, see "Categorías"). In the first place, a decision is made as to whether the predicate is relevant for the annotation of factuality (Applicable / Non applicable). Then, the following aspects are determined: time of the predicate (Present, Past or Future), degree of the writer's commitment to the predicate (Commitment, Non commitment, Qualified commitment), and, finally, polarity (Positive and Negative) and dynamicity (Event, Mental Predicate and Property -Property-Absolute Truth, Eventive Property and Non-eventive Property). At this stage, we also annotate those linguistic marks, either morphological or lexical (triggers) that justify the selected tag. Finally, we also annotate any relevant voices in the narration, other than the writer's, which might modify the view of the event. We can also indicate if there were any problems during the annotation.

4 Conclusions

In this paper we have presented two resources created within the framework of the TAGFACT project: the corpus and the annotation editor. The corpus consists of around 30000 words, created on the basis of 74 pieces of news in Spanish, extracted from three newspapers of different ideologies published in Spain. The extraction was carried out in groups of three pieces of news, each from a different newspaper, covering the same event.

Presently we are manually annotating the pieces of news with regard to the factuality of the events described in them. For this phase of manual annotation, a highly sophisticated editor was created, which makes the annotation easier for the linguist, taking into account that the annotation involves the use of various labels covering the different aspects under consideration.

Both the corpus and the annotation editor are two innovative resources in the field of the annotation of factuality, a task that has become especially relevant in the past few years in NLP. In the future, we intend to contrast the values obtained for the same event in the different newspapers, and implement a system of automatic annotation of factuality based on linguistic knowledge.

Bibliographical references

Diab, M. T., Levin, L., Mitamura, T., Rambow, O., Prabhakaran, V., & Guo, W. (2009). Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop* (pp. 68-73). Singapur.

Lee, K., Artzi, Y., Choi, Y. & Zettlemoyer, L. (2015). Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1643-1648). Lisboa.

Marneffe, M. C., Manning, C. D. & Potts, C. (2012). Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2), 301-333.

Matsuyoshi, S., Eguchi, M., Sao, C., Murakami, K., Inui, K., & Matsumoto, Y. (2010). Annotating event mentions in text with modality, focus and source information. In *Proceedings of the International Conference on Language Resources and Evaluation*, (pp. 1456-1463). Malta.

Minard, A. L., Speranza, M., & Caselli, T. (2016). The EVALITA 2016 Event Factuality Annotation Task (FactA). In *Proceedings of the Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop* (pp.32-39). Naples.

Morante, R., Van Asch, V. & Daelemans, W. (2010). Extraction of biomedical events. *LOT Occasional Series*, 16, 91-105.

Narita, K., Mizuno, J., & Inui, K. (2013). A lexicon-based investigation of research issues in Japanese factuality analysis. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing* (pp. 587-595). Nagoya.

Nawaz, R., Thompson, P. & Ananiadou, S. (2010). Evaluating a meta-knowledge annotation scheme for bio-events. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing* (pp. 69-77). Uppsala.

Padró, L. & Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference*. (pp. 2473-2479). Istanbul.

Saurí, R. (2008). *A Factuality Profiler for Eventualities in Text*. (PhD Thesis). Brandeis Univ.

Saurí, R., & Pustejovsky, J. (2009). FactBank: A corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3), 227-268.

Soni, S., Mitra, T., Gilbert, E., & Eisenstein, J. (2014). Modeling factuality judgments in social media text. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Volume 2: Short Papers* (pp. 415-420). Baltimore.

Van Son, C., van Erp, M., Fokkens, A. & Vossen, P. (2014). Hope and fear: Interpreting perspectives by integrating sentiment and event factuality. In *Proceedings of the 9th International Conference on Language Resources and Evaluation.* (pp. 26-31). Reykjavik.

Velupillai, S. (2011). Automatic classification of factuality levels. A case study on Swedish diagnoses and the impact of local context. In A. Moen, S. K. Andersen, J. Aarts & P. Hurlen (Eds.), In *Proceedings of the XXIII International Conference of the European Federation for Medical Informatics. User Centred Networked Health Care* (pp. 559-563), Oslo.

Wonsever, D., Rosá, A. & Malcuori, M. (2016). Factuality annotation and learning in Spanish texts. In *Proceedings of Language Resources and Evaluation* (pp. 2076-2080). Portoroz.

L'émergence du marqueur méta-discursif *du coup* De la conséquence à l'actualisation énonciative

Lotfi Abouda
Laboratoire Ligérien de Linguistique (LLL, UMR 7270), Université d'Orléans
lotfi.abouda@univ-orleans.fr

Mots-clés : *du coup*, pragmaticalisation, marqueur discursif, français oral hexagonal

1 Introduction

Le peu d'études linguistiques consacrées à l'expression adverbiale *du coup* se sont essentiellement fondées sur l'examen d'exemples forgés (Cadiot & Lebas, 2005, Jayez & Rossari, 1999, Jayez, 2000) ou littéraires (Malm, 2011). Si ces études ont permis de décrire le fonctionnement sémantico-pragmatique de *du coup*, notamment en l'opposant à d'autres marqueurs comme donc ou pour le coup, on peut se demander si les descriptions avancées restent d'actualité dans un domaine, celui des marqueurs discursifs, qui connait une perpétuelle variation. Une telle variation, qui pourrait affecter les valeurs sémantico-pragmatiques de l'expression, est nettement visible au niveau quantitatif, comme l'attestent les nombreux signalements dont elle a fait récemment l'objet de la part de remarqueurs naïfs (Le Figaro 2017, Le Télégramme 2015, Glad 2018). Or le signalement par les locuteurs eux-mêmes d'un usage vu comme singulier et/ou exagéré constitue, selon Siouffi (2012) qu'il nomme événement méta, un mode de repérage utile des micro-variations, et un indice possible de leur caractère innovant.

2 Corpus et méthodologie

Il nous semble ainsi légitime de chercher à dresser une cartographie de l'usage réel de cette expression, et de mesurer l'essor qu'elle a connu en français hexagonal au cours des dernières années.

2.1 Corpus

Pour ce faire, nous nous proposons dans cette étude d'examiner le fonctionnement de ce marqueur discursif, dans le cadre d'une exploration exhaustive et outillée d'un corpus oral, extrait des Enquêtes Socio-Linguistiques à Orléans (ESLO). De taille conséquente (environ 7 millions de mots), ce corpus se distingue par deux caractéristiques essentielles qui rendent faisable cette étude : d'une part, les données ont été collectées, en deux temps, à 40 ans d'intervalle (ESLO 1 entre 1968 et 1971, et ESLO 2 depuis 2010), ce qui rend possible une comparaison micro-diachronique, et, d'autre part, il contient des métadonnées qui spécifient pour chaque enregistrement ses paramètres diastratiques (i.e. les témoins des entretiens sont répartis en tranches d'âge, sexes et catégories socio-professionnelles) et diaphasiques (différents modules, représentant différentes situations communicationnelles, complètent les entretiens : repas, conférences, mais aussi, dans ESLO 2, d'autres modules : école, boulangerie, module 24h ...). Le sous-corpus de cette étude contient 120 heures d'enregistrements, soit environ 1,5 million de mots, et a été

constitué en deux étapes, présentant chacune une coupe particulière des données en fonction de l'hypothèse projetée, micro-diachronique d'abord, synchronique par la suite.

2.2 Méthodologie

Dans un premier temps, nous avons effectué des requêtes sur un corpus micro-diachronique d'environ 1 million de mots (Abouda & Skrovec 2018), contenant des données en tout point comparables prélevées à parts égales dans ESLO 1 et ESLO 2. Par la suite, pour des raisons qui seront explicitées, nous avons dû intégrer de nouvelles données issues d'ESLO 2, ce qui nous a permis d'obtenir un nouveau corpus synchronique d'environ un million de mots présentant une meilleure couverture diaphasique (intégration des « entretiens jeunes », du module 24h ... cf. Abouda & Rendulic 2017).

Extraites grâce au logiciel d'analyse textométrique TXM, les 603 occurrences de l'expression du coup (dont 5 appartenant à ESLO ...) ont fait l'objet d'une annotation affinée visant à préciser leurs propriétés distributionnelles (position dans la phrase, fonction syntaxique) et sémantico-pragmatiques (valeur sémantique, possibilité de substitution par donc).

3 Résultats

Les listes d'occurrences et les étiquettes d'annotation ajoutées ont par la suite été réinjectées sous TXM afin d'être exploitées en textométrie par un croisement des approches quantitative et qualitative. La mise en perspective avec les métadonnées des données statistiques dégagées permet d'une part de retracer l'émergence en micro-diachronie de cette locution en tant que marqueur discursif, et, d'autre part, de vérifier si la fréquence de tel ou tel emploi de cette séquence est tributaire du genre interactionnel et sensible aux variables des locuteurs. L'examen parallèle des variables internes rendra aisée une comparaison systématique de la spécificité sémantique et pragmatico-interactionnelle de chacune des principales valeurs de l'expression *du coup*, qui semble glisser de l'expression de la conséquence à celle de l'actualisation énonciative.

Références bibliographiques

Abouda, L. & Baude, O. (2007). Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des Eslo, in F. Rastier et M. Ballabriga (dir.), *Corpus en Lettres et Sciences sociales. Des documents numériques à l'interprétation*, Actes du XXVIIe Colloque d'Albi, pp. 161-168.

Abouda, L. & Rendulic, N. (2017). Séquence d'introduction de discours représenté : faire ou dire ? *Discours* [En ligne], 21 | 2017, mis en ligne le 22 décembre 2017. URL : `http://journals.openedition.org/discours/9353`

Abouda, L. & Skrovec, M. (2018). Pour une micro-diachronie de l'oral : le corpus ESLO-MD. *SHS Web Conf*, volume 46, 6^{e} Congrès Mondial de Linguistique Française.

Anscombre, J.-C. (éd.). (2009). *Langue française*, 161 : « Les marqueurs d'attitude énonciative », Paris : Larousse.

Breau, A. (2013). *Je dis ça, je dis rien et 200 autres expressions insupportables.* Paris : Leduc.S.

Cadiot, P. & Lebas, F. (2005). Pragmatics of prepositions : A study of the French connectives *pour le coup* and *du coup*, in Kurzon, Dennis & Adler, Silvia (éds), 2008, *Adpositions. Pragmatic, semantic and syntactic* perspectives, 115-132. Amsterdam/Philadelphia : John Benjamins.

Charolles, M. & Vigier, D. (2005). Les adverbiaux en position préverbale : portée cadrative et organisation des discours. *Langue française*, 148, 9-30.

Dostie, G. (2004). *Pragmaticalisation et marqueurs discursifs. Analyse sémantique et traitement lexicographique.* Bruxelles : De Boeck et Duculot.

Glad, V. (2018). Faut qu'on arrête de dire tout le temps « du coup », on fout la honte. *Brain Magazine*, Jeudi 22 mars 2018.

Jayez, J. & Rossari, C. (1999). Du coup. Un connecteur situationnel, in Verschueren, Jef (éd.), Pragmatics in 1998. *Selected Papers from the 6th international Pragmatics Conference*, Volume II, 290-298. Anvers : Ipra.

Jayez, J. (2000). Du coup et les connecteurs de conséquence dans une perspective dynamique. *Lingvisticae Investigationes*, XXIII, 1, 303-326.

Malm, K. (2011). *Une étude de l'expression adverbiale du coup.* Mastergradsoppgave i fransk språk, Fakultet for humaniora, samfunnsvitenskap og lærerutdanning, Universitetet i Tromsø, Våren 2011.

Heiden, S. & al. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. *Actes du 10th International Conference on the Statistical Analysis of Textual Data*, 1021-1032. Rome : Edizioni Universitarie di Lettere Economia Diritto 2.

Heiden, S. (2010). The TXM Platform : Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. *24th Pacific Asia Conference on Language, Information and Computation*, 389-398. Sendai, Japan : Waseda Uiversity.

Siouffi, G. & al. (2012). Comment enquêter sur des diachronies courtes et contemporaines ? *Actes du CMLF*, `http://dx.doi.org/10.1051/shsconf/20120100214`.

La pragmaticalisation de *après* à l'oral : une approche micro-diachronique

Hisae Akihiro [1], Layal Kanaan-Caillol [2] et Marie Skrovec [2].
[1]TUFS Tokyo University for Foreign Studies
[2]Laboratoire Ligérien de Linguistique (LLL, UMR 7270), Université d'Orléans

Mots-clés : Après, Marqueur discursif, Pragmaticalisation, Micro-diachronie, ESLO, Oral

1 Résumé :

De récentes études se sont intéressées aux emplois pragmaticalisés de *après*. Ainsi, tandis que D. Amio et W. De Mulder (2015) examinent l'évolution progressive vers des emplois multicatégoriels et polysémiques en français médiéval dans une perspective diachronique, A. Le Draoulec (2017) s'intéresse aux emplois en synchronie du français contemporain et s'attarde sur celui de connecteur adversatif.

Pour cette étude, nous nous fondons sur la typologie dressée par H. Akihiro (2018), à partir des occurrences identifiées dans le corpus TUFS, un recueil de données orales de style informel, et sur le constat de la forte présence des emplois de *après* en tant que marqueur discursif dans ce corpus.

Nous nous proposons d'examiner la variation des usages de *après* à l'oral à partir de l'annotation d'occurrences dans un corpus structuré sélectionné dans les Enquêtes Sociolinguistiques à Orléans (ESLO) et qui permet d'adopter dans l'analyse un point de vue diachronique (données récoltées à 40 ans d'intervalle, exploitable au sein de la collection ESLO-MD, Abouda & Skrovec 2018) et diaphasique (enregistrements effectués dans des situations présentant des contextes de proximité et de distance).

Une première phase d'annotation permet de consolider l'hypothèse que l'emploi massif de *après* comme marqueur discursif de contraste est relativement récent. Même dans un contexte de proximité communicative, plus susceptible a priori de favoriser l'émergence d'innovations, l'emploi moderne est encore peu fréquent à la fin des années 60. Cependant, on relève déjà quelques emplois macro-syntaxiques de *après* comme articulateur au sémantisme complexe, qui montrent déjà son inscription dans un processus de pragmaticalisation. Ainsi, l'examen systématique, à la fois qualitatif et quantitatif, des occurrences de *après* en micro-diachronie dans le corpus, et la mise en perspective de la typologie adoptée avec les facteurs de variation dits *externes* (Labov 2001) de contexte interactionnel et d'âge, permet une meilleure compréhension de sa pragmaticalisation, étudiée comme processus émergeant en micro-diachronie, ainsi qu'une analyse plus précise de ses emplois contemporains en interaction.

2 Proposition :

Dans des études diachroniques telles que D. Amiot et W. De Mulder (2015) et B. Fagard (2003), il a été constaté que la préposition *après* a connu une évolution progressive jusqu'à devenir un marqueur multi-catégoriel et polysémique. Ces particularités s'observent en français contemporain, comme on le voit dans la description du dictionnaire TLFi. Cependant, certains emplois discursifs, comme par exemple celui de connecteur fréquemment observé dans les conversations, n'apparaît pas dans le TLFi. En se basant sur des exemples issus de Frantext et d'Internet, A. Le Draoulec (2017) décrit les particularités de l'emploi de *après* en tant que connecteur adversatif tout en le comparant avec *maintenant*. Elle remarque également que cet emploi de *après* relève de la pragmaticalisation (cf. Dostie 2004) et que ce processus, proche de la grammaticalisation (cf. Hopper & Traugott 1993, Traugott 1995), ne serait pas encore achevé. En s'appuyant sur des exemples réellement attestés dans le corpus de TUFS, un recueil de données orales de style informel, H. Akihiro (2018) montre qu'il y a, dans l'oral informel actuel, une forte présence de l'emploi de *après* en tant que marqueur discursif.

Jusqu'à présent, aucun travail systématique n'a cependant pu être entrepris sur des grands corpus oraux en vue de mener une analyse à la fois qualitative et quantitative sur l'évolution de l'emploi de cette forme. Si pour A. Le Draoulec (2017), l'emploi relativement récent de *après* comme marqueur pragmatique de "rupture" semble absent de l'écrit alors qu'il prolifère dans les conversations spontanées, on sait peu de choses quant à son émergence à l'oral d'un point de vue diachronique. Nous proposons donc ici, en nous fondant sur la typologie proposée par H. Akihiro (2018), d'examiner la variation des emplois de *après* à l'oral d'un point de vue diaphasique et diachronique, à partir de l'annotation de la totalité des emplois de *après* dans un corpus structuré, sélectionné dans les Enquêtes Sociolinguistiques à Orléans (ESLO).

Ce corpus, en plus de fournir une grande quantité de données orales transcrites (environ 7 millions de mots), présente plusieurs caractéristiques qui permettent d'observer le comportement discursif de *après* et sa pragmaticalisation en français contemporain. Les données, collectées en deux temps (ESLO 1 entre 1968 et 1971, et ESLO 2 depuis 2008), fournissent d'une part un recul diachronique d'une 40aine d'années ; d'autre part, elles sont assorties de métadonnées qui spécifient pour chaque enregistrement ses paramètres diaphasiques (sont documentées, en plus des entretiens sociolinguistiques, des situations de communication variées aux degrés de formalité divers : repas, conférences, discours, micro-trottoir, étudiants dans leur vie quotidienne, etc.), ainsi que diastratiques (âge, sexe et catégorie socio-professionnelle). Le sous-corpus sur lequel se fonde la présente étude, d'un volume total de 1,4 millions de mots, a été constitué de manière à mettre en lumière le comportement de *après* selon ces différents axes.

Ainsi, un premier ensemble de données, le sous-corpus ESLO-MD (Abouda & Skrovec 2018) permettra d'interroger la progression micro-diachronique de *après*. Constitué en grande partie d'entretiens, il est organisé en deux ensembles de données de même volume (respectivement 500.000 mots issus d'ESLO1 et ESLO2) et équilibré d'un point de vue diastratique (panel de locuteurs équilibré constitué selon des critères d'âge, de sexe et de catégorie socio-professionnelle).

Pour observer la variation diaphasique, nous nous appuyons sur des enregistrements supplémentaires qui illustrent différents degrés du continuum variationnel entre proximité et distance communicative (Koch & Oesterreicher 2001) : outre les repas en famille et des conférences universitaires de ESLO-MD, d'autres enregistrements ont été intégrés au sous-corpus d'étude pour dégager un possible contraste entre les contextes de proximité (repas, 24h dans la vie d'une étudiante) et de distance (conférences, discours, assemblées générales). Enfin, une sélection de données produites par des locuteurs jeunes (<25 ans) pourra par ailleurs être examinée afin d'observer l'effet du paramètre générationnel.

Identifiées grâce au logiciel d'analyse textométrique TXM qui permet de croiser approches qualitative et quantitative (Heiden 2010), les occurrences de *après* font l'objet d'une annotation manuelle destinée à dégager les propriétés syntagmatiques et sémantico-pragmatiques de la forme dans ses contextes d'emploi, de manière à faire émerger les configurations discursives associées à sa pragmaticalisation.

Une première distinction, basée sur l'approche pronominale et ses différents tests (notamment la pronominalisation et le clivage, cf. Blanche-Benveniste et al. 1984) permet de distinguer les emplois (micro)syntaxiques comme prépositions ou adverbes, des emplois macrosyntaxiques comme marqueurs discursifs. Ensuite, c'est la position par rapport au constituant dont il dépend qui est annotée : initiale, médiane ou finale.

Enfin une annotation sémantique fine permet de dégager les différentes valeurs exprimées par *après*. En se fondant sur les études antérieures, notamment celles de H. Akihiro (2018) et A. Le Draoulec (2017), nous avons identifié sept valeurs permettant de rendre compte des différents emplois.

Les trois premières se rattachent à la valeur de postérité exprimée par *après*. Il s'agit de la situer, selon les emplois, au niveau temporel (1), au niveau spatial (2) ou au niveau énonciatif.

(1) BA 725 : j'ai fait le métier de boucher euh peut-être pas à contrecœur
mais enfin j'en avais pas tellement
et puis **après** j'ai appris à l'aimer à aimer mon métier
(ESLO_ENT_001)

(2) NS530 : remarque regarde là les écoles SNCF y en a une là q- **après** le pont
(ESLO1_REPAS_275)

La valeur de postérité énonciative s'accompagne selon les cas de quatre autres valeurs. La première, celle d'addition est illustrée dans l'exemple (3), où *après* marque une énumération :

(3) RL 2 : alors moi j'ai des amis qui sont assez jeunes
ch_CD 2 : hm
RL 2 : le plus jeune de mon ami a l'âge de mon fils
RL 2 : donc euh
ch_CD 2 : d'accord
RL 2 : [toux] et donc euh **après** j'en ai des qui ont trente ans et puis **après** j'en ai des plus vieux mais c'est vrai que moi au niveau c'est pas l'âge qui fait euh l'amitié c'est euh souvent des choses qui se passent comme ça et
(ESLO2_ENT_1002)

La deuxième est la valeur de contraste. Nous avons regroupé différents cas de figure dans cette catégorie : elle correspond selon les occurrences à l'introduction d'un argument en opposition à un autre, ou exprime l'idée d'une réserve des interlocuteurs sur le discours en cours, comme en (4), où la locutrice annonce qu'elle ne voit pas de raison de s'opposer à la poursuite d'études des jeunes, tout en émettant une réserve en cas d'incapacité financière des parents :

(4) NS571 : euh quand on a la possibilité de le faire et que et que ça plaît
je vois pas pourquoi euh on s'arrêterait
c'est-à-dire qu'il faut quand même avoir certains moyens ou enfin
je trouve qu'**après** ça dépend quand même des parents
faut que les parents ils puissent euh pousser les les enfants [...]
(ESLO1_ENT_015)

La troisième valeur est celle de consécution. Dans ces contextes, la postériorité induit un lien de cause à effet, le marqueur peut être paraphrasé par « du coup ». A propos des étudiants de mai 68 :

(5) ça a quand même été un petit peu loin ils ont ils ont un peu forcé la dose mais enfin **après** ils étaient plus maitres des évènements
(ESLO1_ENT_015)

La quatrième valeur est celle dans laquelle l'unité opère au niveau méta-discursif, thématisant ainsi le *dire*. Elle est identifiée dans des contextes qui montrent un recul du locuteur par rapport au discours, souvent contrastif, en train de s'élaborer (insertion possible d'une formule comme "je dis ça mais X" , "cela dit") et où l'on relève souvent d'autres commentaires métadiscursifs, comme ici *je sais pas*, dans un extrait où le locuteur évoque ses goûts alimentaires :

(5) ch_OB1 : sinon euh en en en goût comme ça en en en alimentation et tout c'est
euh vous préférez plutôt quoi comme type de bouffe ?
BV1AMIE : la viande
BV1 : ouais voilà la bah la viande mais **après** je sais pas
des vrais plats enfin j'aime bien les vrais plats
(ESLO1_ENT_1001)

La dimension métadiscursive est ici particulièrement prégnante puisque le locuteur émet une réserve sur le choix de la catégorie suggérée par son amie (pas tant la catégorie alimentaire de la "viande" que celle du mode de préparation, le "vrai plat"). Cet exemple montre par ailleurs que les valeurs identifiées ne sont pas exclusives ; en effet, une grande partie des occurrences de *après*, notamment dans ses emplois de marqueur discursif, en actualisent plusieurs, comme ici, où après à la fois contrastif, additif et métadiscursif, est paraphrasable par "mais aussi et surtout" .

Si le nombre d'occurrences totales dans le corpus sélectionné de 1,4 millions de mots ne parle pas en soi (2450 occurrences de *après*), un premier relevé brut dans les sous-corpus MD1 et MD2, de tailles équivalentes, donne un résultat intéressant : on compte, tous types d'emplois confondus, 377 occurrences de *après* dans le sous-corpus plus ancien contre 1137 dans le sous-corpus plus récent. Ces premiers relevés, dans des genres interactionnels comparables (entretiens, conférences et repas entre proches) laissent donc supposer un changement. Après une première phase d'annotation (600 occurrences environ prises dans différents enregistrements), nos observations permettent de consolider l'hypothèse que l'emploi massif de *après* comme marqueur discursif de contraste (ex. 4) est relativement récent : on observe ainsi de nombreuses occurrences de ces emplois dans ESLO2, alors qu'ils semblent, selon les premières estimations, 2,5 fois moins fréquents dans ESLO1. C'est notamment la concentration importante chez certains locuteurs d'ESLO2 du marqueur employé comme élément métadiscursif à des fins de gestion de l'interaction qui témoigne, dans le corpus récent, d'un processus avancé de pragmaticalisation, alliant augmentation de la fréquence en interaction et élargissement de l'éventail fonctionnel. De plus, il est ainsi intéressant d'observer dans les repas entre proches d'ESLO1 une présence importante (46 sur 51 occurrences) des emplois temporels ou spatiaux standards de niveau micro-syntaxique, prépositionnels (15, dont 6 spatiaux) ou adverbiaux (31, dont 1 spatial). Ainsi, même dans un contexte de proximité communicative, plus susceptible a priori de favoriser l'émergence d'innovations, l'emploi moderne est encore peu fréquent à la fin des années 60, et quand il surgit, c'est en emploi additif (4 occurrences) et plus rarement contrastif (2 occurrences). Cependant, ces emplois macro-syntaxiques de *après* comme articulateur au sémantisme complexe ne sont pas complètement absents, ce qui montre son l'inscription dans un processus de pragmaticalisation dès cette période (ex. 4).

A terme, l'examen systématique, à la fois qualitatif et quantitatif, des occurrences de *après* en micro-diachronie, et la mise en perspective de la typologie adoptée avec les facteurs de variation

dits *externes* (Labov 2001) de contexte interactionnel et d'âge, permettra une meilleure compréhension de sa pragmaticalisation étudiée comme processus émergeant en micro-diachronie ainsi qu'une analyse plus précise de ses emplois contemporains en interaction.

Références

Abouda, L. & Skrovec, M. (2018), « Pour une microdiachronie de l'oral : le corpus ESLO-MD », CMLF 2018.

Akihiro, H. (2018), « Discourse function of après in French informal conversation », *Conference Proceedings of the 4th Asia Pacific Corpus Linguistic Conference*, APCLC. 21-28.

Amiot, D. & De Mulder, W. (2015), « Polycatégorialité et évolution diachronique : les emplois préfixoïdes de après (-) et arrière (-) », *Langue française* 187, 137-155.

Blanche-Benveniste, C., Deulofeu, J., Stefanini, J. & Van den Eynde, K. (1984), *Pronom et Syntaxe, L'approche pronominale et son application en français*, Paris : SEALF.

Le Draoulec, A. (2017), « 'Après moi ce que j'en dis···' L'emploi pragmatique de 'après' », *in* Dostie, G. & Lefeuvre, F. (éds), *Lexique, grammaire, discours Les marqueurs discursifs*, Paris : Honoré Champion : 23-40.

Dostie, G. (2004), *Pragmaticalisation et marqueurs discursifs. Analyse sémantique et traitement lexicographique*, Bruxelles : De Boeck et Duculot.

Fragard, B. (2003), « Après : de l'espace au temps, la sémantique en diachronie », in S. Da Silva, A. Torres & M. Gonçalves (éds) *Linguagem, cultura e cognição*, Bragas : Almedina, 231-246.

Franckel, J.-J. & Paillard, D. (2007), *Grammaire des prépositions*, Tome 1, Paris : Ophrys.

Heiden, S. (2010), « The TXM Platform : Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme », in K. I. Ryo Otoguro (Ed.), 24th Pacific Asia Conference on Language, Information and Computation - PACLIC24, 389-398, Institute for Digital Enhancement of Cognitive Development, Waseda University, Sendai, Japan.

Hopper, P. J. & Traugott, E. C. (1993), *Grammaticalization*. Cambridge : Cambridge University Press.

Koch, P. & Oesterreicher, W. (2001), « Langage parlé et langage écrit », in Holtus G., Metzeltin M., Schmitt Ch. (éds), *Lexikon der Romanistischen Linguistik*, Bd. I/2, Tübingen, Niemeyer, 584-627.

Laboratoire Ligérien de Linguistique - UMR 7270 (LLL) (2018). *ESLO-MD : Enquêtes Socio-Linguistiques à Orléans : Corpus Micro-Diachronie* [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - `www.ortolang.fr`, `https://hdl.handle.net/11403/eslo-md/v1`.

Labov, W. (2001), *Principles of linguistic change, Vol. 2 : Social factors*, Oxford : Blackwell.

Schiffrin, D. (1987), *Discourse markers*. Cambridge : Cambridge University Press.

Traugott, E. (1995), « The role of the development of discourse markers in a theory of grammaticalization », paper presented at ICHL XII, Manchester.

Session 6.A.
Prononciation, prosodie

On the link between L2 learner's vocabulary knowledge and pronunciation accuracy: a corpus-based study

Paolo Mairano [1] et Fabian Santiago [2].
[1]STL UMR 8153, Université de Lille
[2]SFL UMR 7023, Université de Paris 8
paolo.mairano@univ-lille.fr, fabian.santiago-vargas@univ-paris8.fr

1 Introduction

In recent years there has been growing evidence that measures of vocabulary size are good predictors of L2 competence, and vocabulary tests are therefore often used as a quick evaluation of L2 proficiency level (Meara, 2010; Milton, 2013). In effect, vocabulary size has been shown to correlate strongly with reading, writing, and listening skills (Stæhr, 2008); this is of course grounded on the fact that knowledge of a higher number of words is likely to result in a better comprehension of text. However, vocabulary tests may give a smaller indication of learners' speaking skills, and in particular pronunciation accuracy. In effect, evidence about the correlation between vocabulary size and speaking skills is scant. Koizumi and In'nami (2013) report on 9 existing studies, most of which evaluate speaking skills in terms of fluency (speech rate, length of utterances, etc.), never in terms of pronunciation accuracy, foreign accentedness or intelligibility – something that may be due to the lack of standard metrics for assessing any aspect of L2 pronunciation. The only existing studies investigating the relation between vocabulary knowledge and L2 pronunciation are in fact very recent. Uchihara & Saito (2019) found significant correlations of productive vocabulary size with speech rate but not with ratings of accenteness and comprehensibility for Japanese learners of L2 English. Similarly, in our previous study (Mairano & Santiago, forthcoming), we found that a measure of receptive vocabulary size (Dialang vocabulary test) showed low to medium correlations with speech rate but not with ratings of foreign accentedness, nor with acoustic measures of vowels, for Italian learners of L2 French. Additionally, we computed metrics of lexical diversity (vocd-D, MTLD, MTLD-MA, cf. McCarthy & Jarvis, 2010) from learners' productions as an indication of their productive vocabulary size and found that they did not correlate significantly with any pronunciation measure.

The aim of this study is to expand the investigation reported by Mairano & Santiago (forthcoming), by computing learners' lexical profiles and verifying their correlation with various L2 pronunciation metrics. In our previous study, we used lexical diversity metrics to estimate learners' productive vocabulary size, as in many other studies (e.g., Arnold et al., 2018). However, low lexical diversity in learners' productions does not necessarily imply little low vocabulary size, while some authors have suggested that lexical profiles may give a better indication of productive vocabulary size (Laufer & Nation, 1995: Edwards & Collins, 2011). This is because low proficient learners tend to reuse frequent lexical items, while more proficient learners tend to use more infrequent lexical items. We therefore expand our previous analysis by computing learners' lexical profiles and verifying if they correlate with our L2 pronunciation metrics.

2 Corpus and methodology

2.1 Corpus

We used the Italian section of the *ProSeg* corpus (Delais-Roussarie et al., 2018), which includes recordings of 25 Italian learners of L2 French in a university setting. Students of L2 French (21 females and 4 males; B1 to C1 levels) at the University of Turin (Italy) were recorded in a sound-proof booth, thereby guaranteeing high-quality audio, apt for acoustic analysis. All participants signed a consent form and filled a questionnaire gathering information about their acquisition process and other useful sociolinguistic information. They were asked to perform the following tasks:

- a read-aloud task of 8 short passages in French (907 words in total)
- a read-aloud of a longer passage in French
- a picture description task
- a monologue (telling a film/book/holiday)
- Dialang vocabulary test
- a read-aloud task of 8 short passages in Italian

The audio was transcribed orthographically and an automatic transcription was forced-aligned to the signal via *EasyAlign* (Goldman, 2011) and subsequently manually checked on *Praat* (Boersma & Weenink, 2018) for all reading tasks and for the initial 5 minutes of the semi-spontaneous tasks (picture description and monologue), taking care to preserve transcriptions that were as close as possible to the target phonemes.

2.2 Learners' lexical profiles

In order to quantify learners' productions from the point of view of word frequency, morphological complexity and phonological complexity, we analysed the first 5 minutes of each learner's production for the picture description task. After lemmatisation of our orthographic transcription with *TreeTagger* (Schmid, 1995), we computed the following metrics on the list of words and lemmas used by every speaker (with reference to the *Lexique 383* corpus, cf. New et al., 2005):

- percent of words ranking >4000, >3000, >2000 and >1000 in the *Lexique* corpus;
- percent of lemmas ranking >3000, >2000 and >1000 in the *Lexique* corpus;
- percent of words with >2 and >1 morpheme (excl. inflectional suffixes);
- percent of words longer than 10, 8, 6 and 4 phonemes.

2.3 Measures of L2 pronunciation

We used various measures of L2 pronunciation, all presented in detail in our previous study (Mairano & Santiago, forthcoming):

- fluency was evaluated in terms of speech rate (SR, phon/sec incl. pauses), articulation rate (AR, phon/sec excl. pauses) and inverted number of pauses (NP);
- global foreign accent was evaluated via ratings of foreign accentedness (FA) on a 5-point Likert scale provided by 3 native French phoneticians (*ICC* = .89) based on 8 sentences extracted from every learner's productions;
- nasal vowels were evaluated via ratings of nasality for /ɛ̃, ɑ̃, ɔ̃/ on a 5-point Likert scale provided by 3 native French phoneticians (*ICC* = .88, .67, .85 respectively for each vowel) based on 9 words extracted from every learner's productions;
- the degree of distinctness of the problematic /y - u/, /ø - e/, /œ - ɛ/ vowel pairs was evaluated via acoustic distances (D) and Pillai scores (P) (Hall-Lew, 2010), following the approach proposed by Mairano et al. (2019) for L1 English.

3 Results and discussion

Firstly, we analysed the relation among our lexical variables by computing a correlation matrix and plotting it as a Fruchterman-Reingold graph via the *qgraph* package (Epscamp et al., 2012). As shown in figure 1, metrics relating to word frequencies (in green) and lemma frequencies (in blue) are all strongly correlated with each other. Additionally, they correlate (more moderately) with metrics of morphological (in yellow) and phonological (in rose) complexity, but not with measures measures of lexical diversity (in orange) nor with learners' scores for the Dialang test.

Finally, we computed the correlation between pronunciation measures and lexical metrics computed on productions of our learners. As shown in figure 2, correlations are very low and mostly do not reach statistical significance: only measures of phonological complexity correlate at $r > .4$ and significantly with (some) pronunciation metrics, suggesting that good pronouncers tend to use longer words. Instead, correlations involving lexical frequency or morphological complexity are low and non-significant, reflecting our previous results obtained with Dialang scores and metrics of lexical diversity. This may be due to the limited dataset and other methodological limitations outlined in Mairano & Santiago (forthcoming); but, for the moment, we cannot provide any tangible proof of a relation between the acquisition of vocabulary and L1 phonology.

References

Arnold, T., Ballier, N., Gaillat, T. and Lissón, P. (2018). Predicting CEFRL levels in learner English on the basis of metrics and full texts. *Proc. of the CAP conference (Conférence sur l'Apprentissage Automatique)*, arXiv:1806.11099

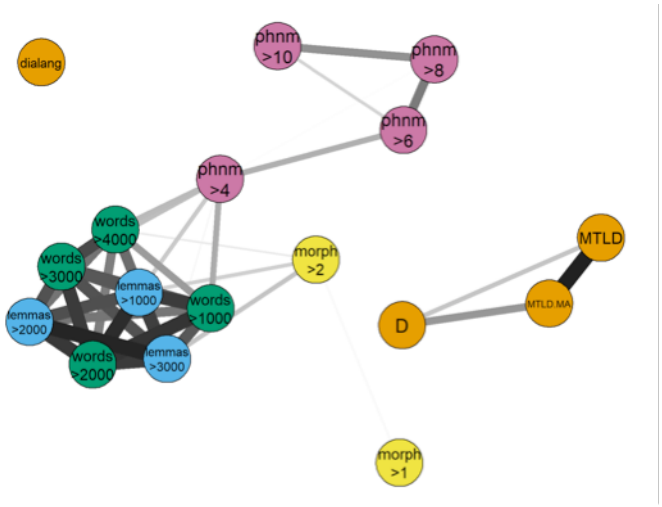


Figure 1: Fruchterman-Reingold graph showing the relation among lexical variables computed on learners' productions.

	FA	D u-y	D e-ø	D ɛ-œ	P u-y	P e-ø	P ɛ-œ	ɛ̃	ɑ̃	ɔ̃	AR	NP	SR
% words with >10 phonemes	0.27	0.51	-0.15	-0.17	0.31	-0.00	-0.03	0.23	0.21	0.29	0.21	0.08	0.17
% words with >8 phonemes	0.09	0.42	-0.22	-0.22	0.28	0.07	-0.22	0.40	0.22	0.04	0.01	-0.13	-0.04
% words with >6 phonemes	0.10	0.33	-0.03	-0.09	0.17	0.32	-0.07	0.17	0.41	0.05	-0.06	-0.06	-0.12
% words with >4 phonemes	-0.36	-0.16	-0.15	-0.21	-0.18	0.03	-0.29	-0.05	0.17	-0.07	-0.28	-0.18	-0.28
% words with >2 morphemes	-0.25	0.04	-0.08	-0.11	0.03	0.26	-0.03	0.01	0.11	-0.05	-0.21	-0.39	-0.25
% words with >1 morphemes	-0.22	-0.18	-0.01	-0.06	-0.14	-0.02	-0.10	-0.01	0.02	-0.16	0.00	-0.31	-0.05
% words with rank > 4000	-0.09	0.01	0.08	-0.02	0.02	0.24	0.13	0.16	0.15	0.01	-0.20	-0.21	-0.18
% words with rank > 3000	-0.27	-0.12	0.20	0.04	-0.12	0.35	0.22	0.02	-0.04	0.03	-0.23	-0.09	-0.23
% words with rank > 2000	-0.14	-0.01	0.34	0.21	-0.00	0.30	0.32	0.06	-0.19	0.19	-0.23	-0.03	-0.21
% words with rank > 1000	-0.45	-0.22	0.14	0.01	-0.22	0.18	0.13	0.01	-0.27	0.02	-0.31	-0.08	-0.27
% lemmas with rank > 3000	-0.18	-0.07	0.24	0.08	-0.10	0.19	0.22	0.15	-0.03	0.29	-0.23	-0.11	-0.22
% lemmas with rank > 2000	-0.07	0.06	0.30	0.15	0.08	0.25	0.31	0.12	-0.21	0.21	-0.17	-0.09	-0.14
% lemmas with rank > 1000	-0.23	0.02	0.14	0.07	0.01	0.36	0.26	0.04	-0.05	0.16	-0.27	-0.21	-0.25

Figure 2: Correlations (Pearson's r) between lexical variables (rows) and pronunciation measures (columns).

Boersma, P. & Weenink, D. (2019). *Praat: doing phonetics by computer* [Computer program]. Version 6.0.49, retrieved 2 March 2019 from `http://www.praat.org/`

Delais-Roussarie, E., Kupisch, T., Mairano, P., Santiago, F. & Splendido, F. (2018) ProSeg: a comparable corpus of spoken L2 French. Poster presented at *EuroSLA*, 5-8 September 2018, Münster (Germany).

Edwards, R., & Collins, L. (2011). Lexical frequency profiles and Zipf's law. *Language Learning*, 61(1), 1-30.

Epskamp, S., Cramer, A. O., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, 48(4), 1-18.

Goldman, J. Ph. (2011). EasyAlign: a friendly automatic phonetic alignment tool under Praat. *Proc. of the 12th INTERSPEECH 2011*, 3233-3236.

Hall-Lew, L. 2010. Improved representation of variance in measures of vowel merger. *Proc. of Meetings on Acoustics* (Vol. 9, No. 1).

Koizumi, R., & In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *Journal of Language Teaching and Research, 4(5)*, 900-913.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics*, 16(3), 307-322

Mairano, P., Bouzon, C., Capliez, M. & De Iacovo, V. (2019). Acoustic distances, Pillai scores and LDA classification scores as metrics of L2 comprehensibility and nativelikeness. *Proceedings of ICPhS2019 (International Congress of Phonetic Sciences)* (pp. 1104-1108), Melbourne (Australia), 5-9 August 2019.

Mairano, P. & Santiago, F. (forthcoming) What vocabulary size tells us about pronunciation skills: Issues in assessing L2 learners. *Journal of French Language Studies.*

McCarthy, P. M. & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods,* 42(2), 381-392.

Meara, P. (2010). *EFL vocabulary tests* (2nd ed.). ERIC Clearinghouse.

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist, & B. Laufer (eds.) *Eurosla Monographs Series*, 2, 57-78.

New, B., Pallier, C., & Ferrand, L. (2005). Manuel de Lexique 3. *Behavior Research Methods, Instruments, & Computers, 36*(3), 516-524.

Richards, B. J. & Malvern, D. (1997). *Quantifying lexical diversity in the study of language development*. Reading: Faculty of Education and Community Studies.

Schmid, H. (1995): Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36(2), 139-152.

Uchihara, T. & Saito, K. (2019). Exploring the relationship between productive vocabulary knowledge and second language oral ability. *The Language Learning Journal*, 47(1), 64-75.

La hiérarchie prosodique affecte-elle l'espace vocalique en français L2 ?

Fabian Santiago [1] et Paolo Mairano [2].
[1]SFL UMR 7023, Université de Paris 8
[2]STL UMR 8153, Université de Lille
fabian.santiago-vargas@univ-paris8.fr, paolo.mairano@univ-lille.fr

1 Introduction

Les voyelles dans les positions prosodiques fortes sont produites avec un plus grand effort articulatoire et se situent dans les positions les plus périphériques de l'espace acoustique vocalique. Ainsi, les consonnes et les voyelles sont plus canoniques (en raison du renforcement de leurs traits phonologiques) lorsqu'elles sont produites dans des syllabes accentuées ou dans des mots associés à une focalisation contrastive prosodique (Cho 2011, Fougeron 2011). Ces segments sont également renforcés au niveau articulatoire lorsqu'ils sont proches de la frontière de certains domaines prosodiques tels que les groupes accentuels (GA) ou les syntagmes intonatifs (SI) en français (Gendrot et al. 2016, Georgeton & Fougeron 2014). Ce phénomène, mieux connu comme *renforcement prosodique*, est corrélé au niveau de la hiérarchie prosodique. Par exemple, en français, plus le domaine prosodique est haut dans la hiérarchie prosodique (syllabe < GA < SI), plus les voyelles sont hyper-articulées et sont plus dispersées dans l'espace vocalique acoustique (Gendrot et al. 2016, Georgeton & Fougeron 2014). De même, en anglais, on observe le même phénomène : les voyelles produites aux positions initiales des *Intonation Phrases* sont plus dispersées dans l'espace vocalique que celles produites dans des positions accentuées (Cho 2005). Cependant, il est difficile de savoir si les effets de la hiérarchie prosodique sur le renforcement prosodique sont observés de manière similaire dans d'autres langues : selon Ortega-Llebaria & Prieto (2007) et Nadeu (2014), la présence/absence d'accents mélodiques (*pitch accents*) et/ou d'accents lexicaux (*lexical stress*) n'est pas un bon prédicteur de la dispersion acoustique des voyelles espagnoles. Nous examinons si la qualité acoustique des voyelles produite par des apprenants adultes de français L2 (anglophones et hispanophones) est affectée par la position prosodique où elles sont produites. En d'autres termes, nous essayons de déterminer si l'expansion de l'espace vocalique reflète la hiérarchie prosodique de la langue cible.

2 Corpus et méthodologie

Nous analysons la parole lue de 30 locuteurs. Les données proviennent de deux corpus : le corpus COREIL (Santiago & Delais-Roussarie 2015) et (ii) le corpus Aix-Ox (Herment et al. 2014). Les analyses ont été faites sur 10 apprenants hispanophones de français L2 (L2FR-ES), 10 apprenants britanniques de français L2 (L2FR-AN) et 10 locuteurs francophones natifs (L1FR). Les participants ont lu neuf courts passages en français décrivant des événements quotidiens (environ 1 minute chacun). L'ensemble des données analysées contient environ 15k voyelles produites. Au moment de la collecte des données, les apprenants suivaient des cours de français de niveau intermédiaire (B1 ou B2) à l'Université Nationale du Mexique (L2FR-ES) et à l'Université d'Oxford (L2FR-AN).

Nous avons identifié les hésitations, les faux départs ou les répétitions qui n'étaient pas annotés originellement dans les données afin de les exclure. Nous avons corrigé la segmentation des phones sous *Praat* (Boersma & Weenink 2016) manuellement. L'analyse finale a été faite sur 12 383 voyelles après filtrage et correction de la segmentation. Nous avons analysé les voyelles produites dans trois positions prosodiques en français (Di Cristo 2010) :

- position finale de syntagme intonatif (SI) ;
- position initiale et finale de groupe accentuel (GA) relevant de l'accent initial (facultatif) et de l'accent final (obligatoire) ;
- position interne de mot (MOT) ou position inaccentuée.

Pour définir ces positions prosodiques, nous avons suivi l'approche employée par Santiago & Delais-Roussarie (2015) afin de faire des comparaisons interlinguistiques robustes entre la parole native et non native. Cela a été réalisé en deux étapes : (i) prédiction de différentes positions prosodiques en fonction de la structure syntaxique, (ii) vérification de ces prédictions sur le signal.

Dans la première étape, la position finale de SI était associée aux frontières droites des phrases coordonnées, des phrases racines et/ou des ajouts en périphérique gauche. Les voyelles produites aux positions initiales du SI ont été exclues de l'analyse car la fréquence de certaines voyelles était déséquilibrée dans nos données, ce qui posait un problème pour le calcul de la surface de l'aire polygonale du triangle vocalique pour certains locuteurs. La position finale du GA était associée à la dernière voyelle de chaque mot lexical, et la position initiale du même GA était associée à la première voyelle des mots lexicaux pluri-syllabiques. Toutes les autres voyelles étaient classées en position interne des mots (MOT), où les voyelles ne sont pas accentuées.

Dans la deuxième étape, nous avons réalisé une analyse acoustique semi-automatique avec le *Prosogramme* (Mertens 2004) afin de corroborer si les voyelles associées aux positions prosodiques ci-dessus étaient produites avec une proéminence prosodique indiquant la frontière des SI et des GA. Ainsi, toutes les voyelles produites avec un mouvement mélodique (descendant, ascendant ou dynamique) couvrant plus de 2 demi-tons avec un seuil de glissando de 0,32 / T2 ont été étiquetées manuellement comme SI (finale) ou GA (position initiale ou finale) en fonction de nos prédictions basées sur la syntaxe. Lorsque ces voyelles n'étaient associées à aucune proéminence acoustique, elles étaient considérées comme voyelles désaccentuées et reclassées dans la catégorie MOT auprès du reste des voyelles en position interne des mots (inaccentuées).

L'ensemble suivant de voyelles orales a été considéré dans l'analyse : / i, e, a, o, u, y, ø, œ /. Un script *Praat* (Boersma & Weenink 2016) a été utilisé pour extraire automatiquement les valeurs des formants F1, F2 et F3 au noyau vocalique afin de minimiser les effets de coarticulation. Les valeurs formantiques ont ensuite été normalisées via l'approche de Lobanov (1971). Pour analyser le renforcement prosodique sur l'espace vocalique, nous avons calculé l'aire du polygone convexe (*Convex Hull Area*, désormais CHA) formée par les valeurs moyennes des formants sur le plan F1xF2 et F2xF3 des voyelles orales mentionnées plus haut par chaque locuteur.

3 Résultats

Les métriques du CHA montrent que les voyelles en français L2 produites dans les positions prosodiques fortes (SI ou GA) forment une aire vocalique acoustique plus large que leurs contreparties inaccentuées dans la position MOT, comme il est montré dans les valeurs reportées à la Table 1. En d'autres termes, le renforcement prosodique est observé en français L2.

	Groupe	Positions prosodiques MOT < GA < SI
F1 x F2	L1FR	2.90 < 3.98 < 5.04
	L2FR-ES	2.70 < 4.50 < 4.94
	L2FR-AN	2.53 < 3.86 < 4.03
F2 x F3	L1FR	1.62 < 2.62 < 5.42
	L2FR-ES	1.23 < 1.45 < 2.52
	L2FR-AN	1.31 < 1.64 > 1.06

Tab. 1 : Valeurs moyennes des aires des polygones convexes (*convex hull areas*) données en écart-types au carré (Lobanov)

Nos métriques montrent également que l'aire de l'espace vocalique est clairement distinguée entre SI et GA en français L1. En effet, l'élargissement de l'espace vocalique sur le plan F1xF2 augmente de 37% de la position MOT à GA, et de 26% de GA à SI. Pour le plan F2xF3, cet élargissement augmente de 61% de la position MOT à GA, et de 106% de GA à SI. Ces résultats montrent que, comme il a été reporté par Gendrot et al. (2016), l'expansion de l'espace vocalique reflète la position prosodique où les voyelles sont réalisées en français L1. En d'autres mots, les locuteurs natifs du français hyper-articulent les voyelles à différents degrés en fonction de la hiérarchie prosodique du français : plus le groupe prosodique est haut dans la hiérarchie, plus les locuteurs ont une tendance à hyper-articuler les segments qui se trouvent proche de la frontière gauche/droite de ces groupes. En français L2, nous observons d'autres scénarios. Le renforcement prosodique en L2 semble avoir un effet plus important entre les positions MOT et GA que celui observé entre GA et SI sur le plan F1xF2. L'élargissement de l'espace vocalique de la position MOT à GA est de 66% pour le groupe hispanophone et de 52% pour le groupe anglophone, alors que chez les natifs est moins important (37%). Plus intéressant encore, l'élargissement de l'espace vocalique entre GA et IP est de 9% pour les apprenants hispanophones et de 4% pour les étudiants anglophones, tandis que chez les natifs ce n'est pas le cas (+26%). Les résultats sur le plan F2xF3 en français L2 montrent aussi des divergences. L'aire de l'espace vocalique s'élargit de 57% de la position MOT à GA, mais seulement 9% de la position GA à SI chez les hispanophones. Chez les anglophones, nous observons un étendu de l'aire de l'espace vocalique de 8% de la position GA à SI, mais une réduction de la position GA à SI de -54%. Ces tendances ne sont pas observées en français L1.

4 Discussion et conclusion

La position prosodique dans laquelle les voyelles sont produites affecte l'espace vocalique en français L2 : les voyelles produites en position prosodique forte (GA ou SI) forment un espace vocalique plus large que celui formé par leurs contreparties en position prosodique faible (MOT). Cela dit, les voyelles en position prosodique forte sont hyper-articulées en L2, tel qu'il est observé en L1. En revanche, l'expansion de l'espace vocalique en L2 n'est pas affectée par la position qu'occupent le domaine prosodiques GA et SI dans la hiérarchie prosodique où ces voyelles sont produites. La taille de l'espace vocalique en L2 formé par les voyelles produites aux frontières des GA et des SI est très similaire (il ne s'élargit pas dans la plupart des cas). Un transfert de la L1 (positif pour les anglophones ou négatif pour les hispanophones) ne peut pas expliquer de tels phénomènes. Nous suggérons que le renforcement prosodique en L2 peut être le résultat d'un degré d'extrême d'hyper-articulation dans toute position prosodique forte (soit GA, soit SI), et ce, indépendamment de la L1 des apprenants.

Références bibliographiques

Boersma, P. & Weenink, D. (2016). *Praat : doing phonetics by computer* (Version 6.0.19).

Cho, T. (2005). Prosodic strengthening and featural enhancement : Evidence from acoustic and articulatory realizations of /a, i/ in English. *J. Acoust. Soc. Am.*, 11.6, 3867–3878.

Di Cristo, D. 2016. *Les musiques du français parlé*. Berlin/Boston : Walter de Gruyter.

Fougeron, C. (2001). Articulatory properties of initial segments in several prosodic constituents in French. *Journal of Phonetics*, 29.2, 109–135.

Gendrot, C., Gerdes, K. & Adda-Decker, M. (2016). Détection automatique d'une hiérarchie prosodique dans un corpus de parole journalistique. *Langue française*, 191.3, 123–149.

Georgeton, L. & Fougeron, C. (2014). Domain-initial strengthening on French vowels and phonological contrasts : Evidence from lip articulation and spectral variation. *Journal of Phonetics*, 44, 83–95.

Herment, S., Tortel, A., Bigi, B., Hirst, D. & Loukina, A. (2014). AixOx, a multi-layered learners' corpus : automatic annotation. In Díaz Pérez, J. & Díaz Negrillo A. (eds). *Specialisation and variation in language corpora*, Bern : Peter Lang, 41–76.

Lobanov, B. M. 1971. Classification of Russian vowels spoken by different listeners. *J. Acoust. Soc. Am.*, 49, 606-08.

Mertens, P. (2004). The Prosogram : Semi-Automatic Transcription of Prosody Based on a Tonal Perception Model. In Bel, B. & Marlien, I. (eds). *Proc. Speech Prosody*, Nara (Japan).

Nadeu, M. (2014). Stress- and speech rate-induced vowel quality variation in Catalan and Spanish. *Journal of Phonetics,* 46, 1–22.

Ortega-Llebaria, M. & Prieto, P. (2007). Disentangling stress from accent in Spanish : Production patterns of the stress contrast in deaccented syllables. In Prieto, P., Mascaró, J. & Solé, M.J. (eds). *Segmental and Prosodic Issues in Romance Phonology*, Amsterdam/Philadelphia : John Benjamins, 155–175.

Santiago, F. & Delais-Roussarie, E. (2015). The acquisition of Question Intonation by Mexican Spanish Learners of French. In Delais-Roussarie, E., Avanzi, M. & Herment, S. (eds). *Prosody and Language in Contact : L2 Acquisition, Attrition and Languages in Multilingual Situations.* Heldelberg : Springer, 243–270.

Session 6.B.
Des corpus pour des études sur l'oral

DECLICS2016 : Un corpus pour recueillir, analyser et améliorer la parole en milieu hospitalier

Mylène Blasco [1], Paul Cappeau [2], Océane Advocat [1], Emmanuèle Auriac-Slusarczyk [3], Aline Delsart [3], Griselda Drouet [4], Yasmine Kebir [5], Elisabeth Richard [4] et Valérie Saint-Dizier de Almeida [5].

[1]LRL, Université Clermont Auvergne

[2]FoReLL, Université Poitiers

[3]ACTé, Université Clermont Auvergne

[4]LIDILE, Université Rennes 2

[5]2LPN, Université de Lorraine

Mylene.Blasco-Dulbecco@uca.fr, paul.cappeau@univ-poitiers.fr

1 Introduction

Notre travail prend place dans un projet de recherche financé trans- et pluri- disciplinaire dans lequel des acteurs sociaux de la santé collaborent avec ceux de la recherche universitaire en sciences humaines (plusieurs domaines de la linguistique, de la psychologie et les sciences de l'éducation). Ce projet analyse des interactions verbales entre soignants et soignés à l'hôpital. Il part d'un constat partagé par tous les acteurs (médecins et patients, personnels de santé) de problèmes dans la communication à l'hôpital. L'intervention de psychanalystes, à la demande de médecins, a ouvert la réflexion sur l'importance plus grande qui devrait être accordée à l'écoute. S'est greffé sur cela le recours à une expertise linguistique en vue d'aider à comprendre la manière selon laquelle les entretiens progressent.

Ce travail centré sur la langue était novateur au regard de nombreux travaux consacrés aux entretiens médicaux (dont Grosjean et Lacoste 1999, Mondada 2006, Ten Have 2006, Ploog et ali. 2018). Il nécessitait de constituer un corpus afin de disposer d'observables fiables dans lesquels les différentes prises de parole pourraient être fidèlement restituées en vue d'analyses.

2 Corpus et méthodologie

2.1 Recueillir une parole sensible

Il n'est probablement pas nécessaire dans le cadre de ces journées de revenir sur des points qui sont communs à la plupart des chercheurs qui travaillent à partir de corpus : l'intérêt de ce type de données est maintenant bien établi (Habert 2000, Cheng 2012, parmi tant d'autres). On préférera ici s'attacher à pointer notamment des spécificités de type éthique que le recueil en situation de santé fait naître (cf. la question du recueil de données langagières dans le cadre de la loi Jardé, publiée le 16/11/2016 et du RGPD applicable depuis le 25/05/2018).

Par les circonstances de recueil et les thèmes abordés, la question de l'anonymisation revêt un caractère plus sensible et délicat que dans d'autres contextes (Baude 2006). Il convient de protéger les patients au-delà de leurs attentes pour être sûr que leur anonymat ne puisse être levé.

Cela a conduit à des décisions concernant la délimitation du cadre institutionnel, la diffusion (en particulier d'extraits sonores), l'anonymisation, le codage et les métadonnées.

2.2 La composition du corpus DECLICS2016

L'une des originalités du corpus repose sur les trois types d'entretiens qui le constituent :

- CO : consultation médicale entre un médecin et un patient,
- EC : entretien clinique entre un psychanalyste et le même patient,
- PC : présentation clinique entre un psychanalyste et un patient face à un auditoire (professionnels de la santé, chercheurs en SHS).

Cette composition permet de disposer de binômes (CO et EC) permettant de contraster ce que dit le patient selon le professionnel avec lequel il interagit. En fonction des analyses développées, d'autres subdivisions du corpus peuvent être exploitées. Ainsi, le service hospitalier peut avoir une incidence forte sur la position du patient : ce dernier est-il en attente d'une désignation de sa maladie (par exemple en neurologie) ou à la recherche d'un aide médicale (par exemple en nutrition) ? De même, le moment où se situe l'enregistrement (première rencontre avec le médecin ou suivi de pathologie) peut-il provoquer (et expliquer) des positionnements différents des locuteurs ?

Ces variables (et d'autres) illustrent la richesse du corpus et les multiples sources d'exploitations qu'il offre.

La transcription est orthographique, choix qui est le plus adapté aux études de morphosyntaxe et aux travaux envisagés.

A ce jour, le corpus se présente ainsi :

Participants	Enregistrements	Transcriptions
37 patients (20 femmes et 17 hommes) 12 médecins (8 femmes et 4 hommes) 6 psychanalystes (2 femmes et 4 hommes)	44 entretiens enregistrés dans les services du CHU 31h 23 min de productions orales Estimation : + 293 400 mots au final 27 CO ; 3 EC ; 14 PC	22 transcriptions vérifiées 15h 23 min de productions orales transcrites 188 214 mots disponibles

Tab. 1 : État d'avancement du corpus DECLICS2016 (juin 2019)

3 Pistes d'exploitation

Le corpus DECLICS2016 bénéficie d'une expérience de constitution de corpus oraux déjà ancienne (le corpus aixois, le CRFP ...) mais il tient son originalité du contexte (le milieu médical, difficile d'accès) et des situations d'enregistrement (pour partie écologiques).

En France, peu d'études ont été réalisées sur les aspects purement linguistiques de l'interaction de soin (Ploog et ali., 2018). Pourtant les recherches axées sur le langage en situation ont sans conteste des choses à dire et à apporter à la formation des soignants. Une double parenté (l'expérience de la linguistique sur corpus et des préoccupations liées aux retombées sociales / sociétales de la linguistique) a orienté vers la constitution d'une équipe pluridisciplinaire pour diversifier les pistes d'exploitation. Ces pistes répondent à des questionnements linguistiques multiples. Chaque spécialiste (lexicologue, syntacticien, psycholinguiste, etc.) impliqué dans le projet apporte son regard, ses outils, sa méthodologie. Les résultats visent à atteindre l'objectif principal : une expertise rigoureuse du corpus à transmettre aux professionnels de la santé. On présentera rapidement trois orientations linguistiques en cours de développement :

a) dans le domaine lexical (Elisabeth Richard et al. 2019), une étude est conduite par le biais de logiciels de traitement de données textuelles, en particulier Iramuteq. Entre calcul fréquentiel, Classification Hiérarchique Descendante (C.H.D.) et analyse factorielle des correspondances (A.F.C.), l'objectif est de saisir ce qui se dit dans ces interactions (de quoi "ça parle"). Les données ainsi quantifiées et hiérarchisées statistiquement permettent d'émettre des hypothèses qui viennent préciser ou différencier le rôle de l'écoute par un psychanalyste ou par un soignant (spécificités, conduite, impact de cette écoute sur le patient, etc.).

b) Dans le domaine morphosyntaxique, on peut se placer dans le cadre de l'approche en « genres » (Biber 2010) et chercher à caractériser le matériau langagier mobilisé par les divers types d'intervenants (patients, médecins, psychanalystes). Il s'agit de s'interroger sur l'éclairage original que ces situations permettent de jeter sur les faits de langue comme par exemple la négation, la densité lexicale ou la forme des groupes nominaux (Advocat & Blasco 2019 ; Blasco & Cappeau 2019 à par.).

c) Dans le domaine pragmatique (Auriac et ali 2019 ; Delsart et ali. 2019, Saint-Dizier de Almeida et ali., 2019), l'objectif est de typifier l'équilibre fonctionnel de l'échange interlocutoire soignants-patient, en posant qu'il existe une variation due au statut différencié entre médecin et psychanalyste. L'étude de certaines marques du discours (pronominalisation, connecteurs discursifs, prise en charge discursives, etc.), comme de certaines opérations discursives (répétitions, reprises, reformulations) situées au sein d'épisodes interlocutoires remarquables servent à repérer des contrastes propres à interroger la formation des médecins (Auriac et ali. Séminaire Acté 2019).

Références bibliographiques

Advocat, O. & Blasco, M. 2019. « Etude d'entretiens médicaux pour parler de densification et de réduction des formes linguistiques à l'oral », In Hana Gruet-Skrabalova et Friederike Spitzl-Dupic (eds) : *Fonctions discursives des formes linguistiques réduites*, Nodus, Münster / Allemagne, 2019. Accepté.

Auriac-Slusarczyk E. & M. Blasco (éds.). 2019. « Les discours des soignants adressés aux patients à l'hôpital. Quelle contribution des sciences humaines et sociales ? ». Revue *ESASO* (Education, Santé, Sociétés). Accepté, en cours d'évaluation pour une publication en juin 2019.

Auriac-Slusarczyk E., Delsart A., Saint-Dizier V., Zehnder E., Blasco M., Advocat O. & Durif F. 2019. « Etude pragmatique des discours soignants/soignés en contexte hospitalier ». Congrès francophone de psychologie de la santé. *Pratiques et interventions en psychologie de la santé*, 13-15 Juin, Metz.

Auriac-Slusarczyk, E., Advocat O., Blasco M., & Delsart A. 2019. « Le projet structurant inter-laboratoires DE-CLICS : des données recueillies à la relation de soin à l'hôpital ». Séminaire du laboratoire Acté, ESPé, Clermont-Ferrand le 8 avril.

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

Biber, D. & Conrad, S. 2010. Register, Genre ans Style, Cambridge : Cambridge University Press.

Blasco, M & Cappeau, P. 2019. (à par.) « Construire et analyser un corpus oral sur objectifs spécifiques : précautions et réflexions ». Publication de la JE "Corpus sur objectifs spécifiques" Lyon3, Université Jean Moulin (15-16 nov. 2018), Del Bove M., Gautier L., Jamet D. et Millot Ph. (eds). En cours d'évaluation.

Cheng, Winnie. 2012. *Exploring Corpus Linguistics. Language in Action.* Abingdon. Routledge.

Delsart, A. et Marquès A. 2019. « Effet de l'expertise communicationnelle des soignants sur la prise de parole des patients. Etude comparative des discours entre médecins et psychanalystes ». Dans E. Auriac-Slusarczyk & M. Blasco (éds.). « Les discours des soignants adressés aux patients à l'hôpital. Quelle contribution des sciences humaines et sociales ? ». Revue *ESASO* (Education, Santé, Sociétés).

Grosjean, M. & Lacoste, M. 1999. Communication et intelligence collective. *Le travail à l'hôpital*. Paris : PUF.

Habert, B. 2000. "Des corpus représentatifs : de quoi, pour quoi, comment ?" dans Bilger, Mireille (éd). *Linguistique sur corpus –Etudes et réflexions.* Perpignan. P.U. de Perpignan. 11-58.

Mondada, L. 2006. Interactions en situations professionnelles et institutionnelles : de l'analyse détaillée aux retombées pratiques. *Revue française de linguistique appliquée*, XI (2), 5-16.

Ploog, K. Mariani-Rousset, S. & Hutin, E. (dir.). 2018. *Emmêler & démêler la parole. Approche pluridisciplinaire de la relation de soin*, PUFC (coll. Annales littéraires).

Richard, E. & Drouet, G. & Moreau, F. 2019. « Analyse lexicométrique du corpus DECLICS : une approche quantitative et qualitative ». Dans E. Auriac-Slusarczyk & M. Blasco (éds.). « Les discours des soignants adressés aux patients à l'hôpital. Quelle contribution des sciences humaines et sociales ? ». Revue *ESASO* (Education, Santé, Sociétés).

Ten Have, P. 2006. On the interactive constitution of medical encounters. *Revue française de linguistique appliquée*, 2 (Vol. XI), 133-162.

Saint-Dizier de Almeida, V., Zehder, E., & Kebir Y. 2019. « Constitution de ressources pour former à la conduite des entretiens médicaux de suivi en CHU ». Dans E. Auriac-Slusarczyk & M. Blasco (éds.). « Les discours des soignants adressés aux patients à l'hôpital. Quelle contribution des sciences humaines et sociales ? ». Revue *ESASO* (Education, Santé, Sociétés).

CIEL-F project: a comparable and ecological corpus to study spoken and interactional practises of spoken French around the world.

Daniel Alcon [1] et Carole Etienne [2].
[1]Department of Romance Languages & Literatures, University of Freiburg (Germany)
[2]ICAR, CNRS / ENS de Lyon / Université Lyon 2
daniel.alcon@romanistik.uni-freiburg.de, carole.etienne@ens-lyon.fr

Keywords: interaction, spoken french, ecology, comparative analysis, tools, TEI

1 Presentation of CIEL-F project

Ciel-F is a corpus of spoken French collected in the francophone community, which gathers around 200 recordings of ten minutes. They were collected from 2006 to 2012 in 15 different regions across the world. This project concerns five teams directed by professors in different universities, each of them responsible of a given region: Lorenza Mondada (Lyon), Françoise Gadet (Paris-Ouest), Stefan Pfänder (Freiburg), Ralph Ludwig (Halle) and Anne-Catherine Simon (Louvain-la-Neuve).

The primary data of the collection was recorded in everyday situations and in professional settings, without any instruction given to the speakers to ensure naturally occurring data. To make things comparable, we chose 3 types of interaction in each area: professional, table-talk and local radio, collected in several towns of the 15 regions (Gadet & al. 2012). This initiative will allow the study of the similarities and the diversities of spoken French produced by different speakers of different regions, since they are involved in the same kind of everyday settings (Mondada & Pfänder 2016).

The choice of the setting is due to the fact that these situations can be found in all of the regions. This choice includes also practical aspects, like a limited number of speakers, in order to avoid noisy recordings; respect of ethical and juridical constraints; research interests, among others.

The choice of regions is not only linked to geographical criteria, but also to the social aspect of French language, French language as a vector, the type of linguistic contact, a living practise of French, ...

The constitution of such a corpus implies a methodological approach in which we have to make decisions during the different stages of the project. For example:

- Choosing types of interactions which we can find in each region
- Recording the data with the same technical configurations and recording tools
- Choosing types of interactions which we can find in each region

- Recording the data with the same technical configurations and recording tools
- Choosing a common but realistic subset of metadata collected simultaneously with the recording of the speakers
- Deciding on a set of transcription conventions and choosing an alphabet to deal with words borrowed from other languages
- Anonymizing both transcripts and recordings with the same procedures and technical solutions
- Fixing the standardized formats for primary and secondary data
- Hosting data in a stable website with good maintenance and search tools in order to make analyses easier for researchers

2 Making corpora comparable and interoperable

In order to study each different corpus in the same way, we need a common subset of metadata for all regions which matches the main information relevant to both compare data and get information during analyses. In fact, researchers working on interaction often use qualitative analyses, looking precisely to usage in context.

Due to calendar constraints, we could not benefit from the work on metadata made in the french CORLI consortium or in the ANR project Orfeo. Nevertheless, our challenges gave nice examples to the CORLI team and we will be able to deliver metadata and data in TEI according to the CORLI consortium recommendations to make them easily reusable in oral European community.

We decided that the corpora will be hosted in both a German and a French database to allow people working usually in one of the platforms to test the tools of the other one and take advantage of both search tools (Alcón, Groupe ICOR 2016). Thus, we decided to define a TEI format to exchange metadata to avoid a double entry which will avoid mistakes and wastes of time. Our solution implements the collection concept in TEI using teiCorpus elements. We don't find many examples in TEI documentation, so we will present it shortly in our presentation.

3 Similarity studies: purposes and proposals

As we have extensive data available —around 48h of data corresponding to 200 transcripts of 200 recordings of at least 10 minutes— we need to organize things to make search tools relevant for interactional studies.

Similarity studies could be particularly interesting in the CIEL-F project for the following purposes:

- Organisations of turns: length (long/short turns), overlaps, pauses ...
- Frequency of tokens, of expressions and of structures
- Co-occurrences
- Oral markers
- Dislocation
- Repetitions by the same speaker or other-repeats
- Some common expressions

We have reflected on the best way to organize results in order to display information in an easy-to-use visualization based on cartography, interaction type or speakers' metadata ... and, of course, on a combination of all of them.

An overview of search tools available on oral corpora underlines the lack of existing tools in order to compare results between different subset of corpora, it seems that users need to execute the same search tool separately on each subset and deal with the comparison of the results without any easy display to bring out the differences or the similarities.

In this project, we will get interested in a way to point out the common features and both the main differences in practises of spoken French all around the world depending on the area (Skrovec & Pfänder 2012), on the setting or on some sociolinguistic criteria. We will for example compare some practises between professional and private settings or between radio and face-to-face interaction. We will need of course quantitative tools to give an idea of the similarities level but we will go back to each excerpt for an analyse in context according to the interactional approach.

In our presentation, we will describe our methodological approach, present the TEI based exchange format concerning metadata and make some proposals on search tools to make possible similarity studies.

References

Alcón, D. (to be published).Moca : Multimodal Oral Corpus Administration.

Gadet, Françoise & Ludwig, Ralph & Mondada, Lorenza & Pfänder, Stefan & Simon, Anne Catherine. (2012) Un grand corpus de français parlé: le CIEL-F. Choix épistémologiques et réalisations empiriques. In *Revue Française de Linguistique Appliquée. Langue parlée: norme et variations* 17(1), 39-54

Mondada, L. & Pfänder, S.(2016) . Corpus international écologique de la langue française (CIEL-F): un corpus pour la recherche comparée sur le français parlé. Corpus n° 15, *Corpus de français parlés et français parlés des corpus*

Pfänder, S. & Skrovec M. (2011). Donc, entre grammaire et discours. Pour une reprise de la recherche sur les universaux de la langue parlée à partir de nouveaux corpus. in M.Drescher & I. Neumann-Holzchuh (éd) *Syntaxe de l'oral dans des variétés non hexagonales du français. Tübingen:Stauffenburg Verlag*

Groupe ICOR (Bert, M., Bruxelles S., Etienne C., Jouin-Chardon E., Lascar J., Mondada L., Teston S. Traverso V.). (2010). Grands corpus et linguistique outillée pour l'étude du français en interaction (plateforme CLAPI et corpus CIEL), Pratiques 147-148 *Interactions et corpus oraux*, 17-35

Session 7.A.
Des corpus autour des écrits scolaires

Premières exploitations textométriques d'un corpus scolaire longitudinal (CP-CM1)

Claude Ponton , Claire Wolfarth et Catherine Brissaud .
Lidilem, Université Grenoble Alpes
claude.ponton@univ-grenoble-alpes.fr, claire.wolfarth@univ-grenoble-alpes.fr,
catherine.brissaud@univ-grenoble-alpes.fr

1 Introduction

Dès 2007 et jusqu'en 2011, Marie-Laure Elalouf et Catherine Boré pointaient le manque de grands corpus scolaires pour appuyer les recherches en didactique du français (Elalouf & Boré 2007 ; Elalouf 2011). Depuis, plusieurs recherches se sont développées avec, entre autres, comme objectif la constitution de tels corpus (David & Doquet 2016 ; Boré & Elalouf 2017 ; Wolfarth *et al.* 2018a). Dans cette même optique, le projet ANR E-Calm[1] vise la constitution d'un large corpus du primaire à l'université dans lequel sont intégrés certains des corpus cités précédemment. Cette communication s'intéressera plus spécifiquement au corpus Scoledit qui est l'un des corpus du projet E-Calm. La spécificité de ce corpus est le suivi de la même cohorte d'enfants du CP au CM2 permettant ainsi une analyse longitudinale des compétences en littéracie au primaire. Après avoir présenté ce corpus, sa constitution et son état d'avancement, nous proposerons une première analyse textométrique de ces écrits selon différentes variables décrivant les écoles et les élèves.

2 Constitution du corpus

2.1 Protocole et recueil

La recherche « Lire-Ecrire au CP » (Goigoux 2016) a permis le recueil de divers écrits (dictées, productions de texte) auprès de plus 2000 enfants suivis du CP au CE1. La recherche Scoledit poursuit cette approche longitudinale sur les niveaux suivants (CE2-CM2) en restreignant toutefois le nombre d'enfants suivis (1135 élèves). Ce second recueil, effectué d'avril 2016 à juillet 2018, a permis de recueillir 1135 productions de texte en classe de CE2 (2016), 1132 productions de texte en classe de CM1 (2017) et 1030 productions de texte en classe de CM2 (2018). Pour diverses raisons (absence d'enfants, impossibilité de suivi de certaines classes ...), le nombre d'enfants pour lesquels nous disposons de l'ensemble des données du CP au CM2 est de 373, répartis dans 31 écoles situées dans les 5 académies suivantes : Clermont-Ferrand, Grenoble, Lyon et Toulouse. C'est cet ensemble, nommé *corpus longitudinal*, qui constitue l'objet de cette communication.

Les productions sont toutes des textes narratifs (il s'agissait de raconter une histoire) avec une consigne spécifique en CP (*ie.* celle de la recherche Lire-Ecrire au CP) et la même consigne (mais avec des temps variables de passation) du CE1 au CM2. L'ensemble de ces productions

1. http://e-calm.huma-num.fr/le-projet/

est assorti de métadonnées décrivant des caractéristiques de l'école (localisation géographique, composition sociale) et des enfants (genre, mois et année de naissance, redoublement ou non au CP, langues parlées à la maison, CSP des parents).

Pour répondre aux objectifs de cette recherche et permettre ainsi une exploitation et un partage libre des données recueillies, chaque production est anonymée et nous disposons des autorisations parentales nécessaires à leur exploitation. L'ensemble du corpus et des traitements visés ont également fait l'objet d'une déclaration préalable auprès de la CNIL[2].

2.2 Numérisation

Chacune des productions a été numérisée dans un quadruple objectif. Premièrement, il s'agit de sauvegarder et de pérenniser les données. Deuxièmement, la version numérique facilite la transcription des copies ; les scans sont directement accessibles via la plateforme de transcription. Le troisième point, central dans la démarche de cette recherche, consiste à partager avec les communautés éducative et de recherche l'ensemble des données recueillies ; actuellement, le corpus est accessible à l'adresse `http://scoledit.org/scoledition`. Enfin, lors des exploitations futures du corpus, ces scans permettent un retour à la copie d'origine qui offre une iconicité maximale (Boré & Elalouf 2017).

2.3 Transcription

Pour permettre l'exploitation informatique du corpus, une étape dite de transcription est nécessaire. Elle consiste, à partir des scans, à produire une représentation textuelle des copies. Le résultat de cette étape constituera la donnée source de tous les traitements et analyses à venir. Il est donc primordial de la maitriser afin d'obtenir des données de qualité répondant aux exigences des recherches à mener. Les réponses aux questions « que transcrire ? » et « comment transcrire ? » ont été décrites dans un guide de transcription éprouvé auprès d'une quinzaine de transcripteurs différents. Globalement, la transcription Scoledit est une transcription de type diplomatique[3] qui permet un affichage aligné entre scan et transcription (cf. fig.1) et dans laquelle seule la production finale de l'enfant est conservée. Ces choix de transcription sont issus d'un travail collaboratif des équipes Clesthia[4], CLEE ERSS[5] et Lidilem[6] repris dans le cadre du projet E-Calm.

2.4 Normalisation

Les transcriptions ne permettent pas d'analyses avancées dans les textes comme rechercher les différentes graphies d'une forme, repérer les temps verbaux ou des structures syntaxiques

2. Commission Nationale de l'Informatique et des Libertés, `https://www.cnil.fr/`

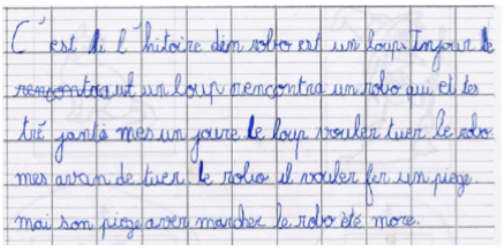
3. « la transcription diplomatique photographie le document en rapportant, avec les outils qui le permettent, malgré leurs limites, tous les événements du manuscrit » (Crasson & Fekete 2007).

4. Université Paris3, EA 7345.

5. Université Toulouse Jean Jaurès, UMR 5263.

6. Université Grenoble Alpes, EA 609.

précises, etc. L'utilisation d'outils classiques de textométrie ou d'outils issus du TAL n'est également pas possible au vu du fort éloignement à la norme de ces productions. Les analyses outillées requièrent donc à minima une version dans laquelle les erreurs de segmentation en mots (hyper/hypo segmentation ; exemple en rouge fig.1) et d'orthographe soient corrigées. Nous avons fait le choix de produire manuellement cette deuxième version appelée « normalisation ». Cette version peut être vue comme une annotation déportée, indépendante de la transcription, à l'instar des projets sur les SMS de Fairon *et al.* (2007). L'idée principale de la normalisation est de rester au plus près de la production de l'enfant pour permettre des aller-retours entre transcription et normalisation à l'aide d'un outil d'alignement production/normalisation développé dans l'équipe (Wolfarth *et al.* 2018b). Comme pour l'étape de transcription, des choix de normalisation ont été faits et sont décrits dans un guide de normalisation.



C'est [x] l'histoire din robo est un loup. Injour [x]
[x] un loup rencontra un robo qui et tes
tré jant[x]i mes un joure. Le loup vouler tuer le robo.
mes avan de tuer. Le robo il vouler fer un piege.
mai son piege aver marcher le robo été more.

C'est l'histoire d'un robot et un loup. Un jour un loup rencontra un robot qui était très gentil mais un jour le loup voulait tuer le robot. Mais avant de tuer le robot il voulait faire un piège. Mais son piège avait marché <s/> le robot était mort.

FIG. 1 : Alignement scan/transcription puis normalisation, élève 200, CE

3 Analyse textométrique

Les transcriptions et les normalisations constituent le matériau de départ des analyses à venir. La qualité de ces données est donc primordiale. Pour cela, chaque transcription et normalisation sont revues et validées par les chercheurs de l'équipe Scoledit. Actuellement, le corpus longitudinal est entièrement recueilli et le travail de scan/transcription, de normalisation et de validation est largement avancé. Pour cette communication, l'intégralité des productions CP-CM1 sont transcrites et normalisées. Le travail sur le CM2 étant en cours, il ne sera pas traité dans le cadre de cette communication qui portera donc au final sur 4x373 textes (soit 1492 textes transcrits et normalisés).

Initialement la plupart des variables d'analyse (comme celles portant sur l'enseignant) définies lors de la recherche « Lire-écrire au CP » ne concernent que le niveau CP. Pour notre propre étude, nous n'avons retenu que celles ayant une portée longitudinale à savoir :

- Écoles : ville, académie, composition sociale : école prioritaire (EP), défavorisée (DF), favorisée (FV), mixte (MX)

- Élèves : sexe, mois et année de naissance, redoublement ou non au CP (RCP), langues parlées à la maison (1 : français, 2 : autre, 3 : français et autre), CSP des parents (de 1 à 8).

Notre corpus longitudinal compte actuellement 373 enfants (214 filles, 159 garçons) suivis sur les niveaux de CP à CM1 et répartis dans 31 écoles appartenant aux académies de Clermont-Ferrand (13 écoles), Grenoble (11 écoles), Lyon (5 écoles) et Toulouse (2 écoles).

La composition de chacune de ces 31 écoles est la suivante :

- Mixte : 16 écoles, 225 élèves
- Éducation prioritaire : 7 écoles, 46 élèves
- Défavorisée : 6 écoles, 70 élèves
- Favorisée : 2 écoles, 32 élèves

Développé dans l'ANR Textométrie[7], TXM est une plateforme ouverte dédiée à l'analyse de grands corpus textuels (Heiden 2010). Couplé au logiciel d'étiquetage morphosyntaxique TreeTagger[8] (Schmid 1994), il permet d'étudier les répartitions lexicales, les cooccurrences, les vocabulaires spécifiques... au regard des différentes variables énoncées ci-avant. Nous proposerons lors de la communication une approche plutôt inductive de type corpus-based (Teubert 2009) pour explorer l'impact des variables étudiées (niveaux, enfants, écoles) sur les caractéristiques de notre corpus.

Références bibliographiques

Boré C., Elalouf M.-L. (2017). « Deux étapes dans la construction de corpus scolaires : problèmes récurrents et perspectives nouvelles ». In C. Doquet, J. David, S. Fleury, Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement. *Corpus*. 16 | 2017.

Crasson A., Fekete J.D. (2007). Structuration des manuscrits : Du corpus à la région. Item [version en ligne : `http://www.item.ens.fr/index.php?id=173027`].

David J., Doquet C. (2016). « Les écrits d'élèves : un corpus de référence pour le français contemporain ». *Congrès Mondial de Linguistique Française*. Juillet 2016. Tours. France.

Elalouf M.-L. (2011). « Constitution de corpus scolaires et universitaires, vers un changement d'échelle ? ». *Pratiques*. n°149-50.

Elalouf M.-L., Boré C. (2007). « Construction et exploitation de corpus d'écrits scolaires ». *Revue française de linguistique appliquée.* vol. xii. no. 1. pp. 53-70.

Fairon C., Klein J. R., Paumier S. (2007). *Le langage SMS : étude d'un corpus informatisé à partir de l'enquête « Faites don de vos SMS à la science ».* Presses univ. de Louvain. Belgique.

Goigoux R. (2016). *Lire et Écrire. Étude de l'influence des pratiques d'enseignement de la lecture et de l'écriture sur la qualité des premiers apprentissages.* Institut Français de l'Éducation. Rapport de recherche.

7. ANR-06-CORP-029, 2007-2010, `http://textometrie.ens-lyon.fr/`
8. `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`

Heiden S. (2010). « TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement ». In *Proceedings of 10th International Conference on the Statistical Analysis of Textual Data.* JADT 2010. Juin 2010. Rome. Italie

Schmid H. (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees" . In *Proceedings of International Conference on New Methods in Language Processing.* Manchester. UK.

Teubert W. (2009). « Corpus Linguistics : An Alternative ». *Semen*. 27 | 2009

Wolfarth C., Brissaud C., Ponton C. (2018a). « Transcrire et normer un corpus scolaire : pour quelles analyses ? ». *Dyptique*. 36. Presses Universitaires de Namur

Wolfarth C., Ponton C., Brissaud C. (2018b). "*Which Method to Develop a Natural Language processing Tool to automatically analyze First Language Learner Corpora ?*" . In *3th Teaching and Language Corpora Conference* (TALC2018). 18-21 July 2018. Faculty of Education. University of Cambridge

Session 7.B.
Autour des genres textuels

Phraséologie et genres textuels : étude des phraséologismes construits autour des verbes *doner* et *metre* dans le roman médiéval

Corinne Denoyelle [1,2] et Julie Sorba [2,1].
[1]Litt&Arts UMR 5316, Université Grenoble Alpes
[2]Lidilem, Université Grenoble Alpes
Corinne.Denoyelle@univ-grenoble-alpes.fr, Julie.Sorba@univ-grenoble-alpes.fr

1 Introduction

Notre contribution s'inscrit dans le cadre des études en phraséologie et s'appuie sur l'exploration outillée d'un corpus littéraire de langue française médiévale. Dans la lignée de Legallois & Tutin (2013), nous considérons la phraséologie dans sa conception étendue. En effet, nous postulons que les unités repérées par des moyens statistiques occupent une fonction structurante dans le texte littéraire (Siepmann 2015 & 2016). Notre objectif est donc de montrer comment ces unités permettent d'organiser les séquences discursives au sein des textes littéraires médiévaux qui constituent notre corpus.

2 Corpus et méthodologie

2.1 Corpus

Notre corpus est composé de romans en prose du XIII^{e} siècle (*Tristan, Lancelot, Artus de Bretagne, La Queste del Saint Graal*) que nous contrastons avec des œuvres de la même période de genres textuels différents : romans en vers (romans idylliques, *Roman de Silence*, romans arthuriens) et chroniques (*Conquête de Constantinople, Histoire ancienne jusqu'à César*).

2.2 Méthodologie

Notre exploration outillée s'appuie, d'une part, sur l'interface LGerM (`http://www.atilf.fr/LGeRM/`), et, d'autre part, sur le Lexicoscope, un outil d'extraction des séquences phraséologiques basé sur des corpus arborés (`http://phraseotext.univ-grenoble-alpes.fr/lexicoscope/`, Kraif 2016). À partir des données extraites, nous procédons à un retour aux textes pour assurer nos analyses. Celles-ci s'inscrivent dans le cadre des approches fonctionnalistes et contextualistes qui intègrent la dimension discursive à l'analyse syntaxique et sémantique des unités linguistiques (Sinclair 2004).

3 Résultats

À partir de l'observation des unités phraséologiques construites autour de deux verbes pivot *doner* et *metre*, nous faisons l'hypothèse que les constructions lexico-syntaxiques récurrentes permettent de contraster différents genres textuels. En effet, de premières observations montrent que le collocatif nominal privilégié qui se construit avec *doner* varie selon les genres : les romans en prose favorisent la construction *doner* DET *coup* dans son acception guerrière (ex. 1) alors que

ce sont des richesses, de l'argent, des épouses que l'on donne comme biens dans les chroniques (ex. 2).

(1) et li **done** si grant **cop** qu'il le porte dou cheval a terre. (*Tristan*, §674)

(2) et disent qu'il n'en renderoient mie sans grant avoir tant que li message leur **donnerent tant d'or et d'argent** comme il demanderent. (*Conquête*, XXVI,8)

De même, le verbe *metre* se rencontre dans les romans en prose plutôt en compagnie de collocatifs nominaux relevant du champ sémantique du combat (*épée, écu, à mort* etc., ex. 3) alors que des noms plus abstraits semblent plus spécifiques aux autres genres textuels (*metre sur quelqu'un « accuser », metre en avant quelque chose, metre conseil, « donner la parole », metre les siens, « placer ses hommes »* etc., ex. 4).

(3) si **met le glaive** sos l'aissele et fiert le cheval. (*Lancelot*, 8, 31)

(4) Quant li marchis oï chou si i vaut **metre les siens** et chiax que il cuidoit qui l'esleussent a empereur. (*Conquête*, XCV.11)

L'étude propose une étude des variations syntagmatiques et paradigmatiques de ces constructions afin de mettre en évidence leurs spécificités dans les différents genres textuels.

Références bibliographiques

Chaurand, J. (1963). Les verbes supports en ancien français : *donner* dans les œuvres de Chrétien de Troyes. *Linguisticae Investigationes* VII/1, 11-46.

Kraif, O. (2016). Le lexicoscope : un outil d'extraction des séquences phraséologiques basé sur des corpus arborés. *Cahiers de lexicologie*, 108, 91-106.

Legallois, D., Tutin, A. (2013). Présentation : vers une extension du domaine de la phraséologie. *Langages*, 189, 3-25.

Marchello-Nizia, C. (1996). Les verbes supports en diachronie. *Langages*, 121, 91-98.

Siepmann, D. (2015). A corpus-based investigation into key words and key patterns in post-war fiction. *Functions of language*. 22(3), 362-399.

Siepmann, D. (2016). Lexicologie et phraséologie du roman contemporain : quelques pistes pour le français et l'anglais. *Cahiers de lexicologie*, 108, 21-41.

Sinclair, J. (2004). *Trust the Text : Language, Corpus and Discourse.* Londres : Routledge.

Posters

Degré d'implication du scripteur dans les textes argumentatifs produits par des apprenants sinophones du français

Tatiana Aleksandrova [1] et Catherine David [2].
[1]LIDILEM, Université Grenoble Alpes
[2]LPL, Université Aix-Marseille
tatiana.aleksandrova@univ-grenoble-alpes.fr, catherine.david2@univ-amu.fr

1 Introduction

Cette étude s'intéresse aux productions écrites d'apprenants sinophones du français afin d'identifier les stratégies d'organisation textuelle qu'ils mettent en place et notamment le degré d'implication du scripteur dans ces textes. De nombreux travaux dans le domaine de l'acquisition des langues secondes (L2) s'intéressent aux processus d'acquisition et aux difficultés d'apprenants de différentes langues premières (L1) à acquérir les compétences écrites en L2. Ils montrent que la distance entre les langues représente un critère important. Par exemple, les productions d'apprenants asiatiques en anglais et en français L2 sont souvent jugées incohérentes par les locuteurs natifs (Bi, 2016).

Nous nous appuyons sur le modèle proposé par Wang & Wen (2002) qui présente le processus d'écriture sous forme de trois composantes : l'environnement de la tâche, le processus de composition et la mémoire à long terme. Ces éléments se trouvent en interaction permanente et sont plus au moins attachés à la L1 du scripteur. Hidden (2012) montre que suite à la surcharge cognitive, les apprenants de différentes L1 se focalisent systématiquement sur la forme de la phrase et font moins attention à l'organisation du texte. Les travaux en rhétorique contrastive évoquent cinq critères d'analyse textuelle permettant de dégager des tendances propres aux scripteurs de différentes langues (Connor, 1996). Ces critères sont : le plan, la linéarité, la connexité, le style et le degré d'implication du scripteur. Dans notre étude, nous focalisons notre attention sur ce dernier critère car nous pensons que les différences culturelles et linguistiques influencent l'engagement du scripteur dans le texte.

La culture chinoise basée sur le confucianisme cherche à harmoniser les relations (Bi, 2016). Les scripteurs doivent exprimer leur respect et la politesse envers le destinataire. La critique et l'expression des émotions négatives sous une forme agressive peuvent avoir une atteinte à la face du destinataire. Ils essaient donc d'éviter ce genre de comportement. En revanche, il est tout à fait possible d'exprimer son désaccord et son mécontentement en France. Les scripteurs ont moins de retenue.

Du point de vue linguistique, les deux langues offrent aux scripteurs des moyens pour marquer la présence de l'auteur comme les pronoms personnels et les adjectifs possessifs. En revanche, leur statut n'est pas le même. En français, le pronom sujet est obligatoire dans la phrase, alors qu'en chinois il peut être omis. Quant aux adjectifs possessifs, ils sont également obligatoires en français et facultatifs en chinois, langue dans laquelle les déterminants ne sont pas obligatoires.

Les formes impératives, qui permettent également de marquer l'engagement fort du scripteur, sont disponibles dans les deux langues. Nous cherchons donc à savoir comment les apprenants issus d'une culture éloignée et maitrisant une langue typologiquement différente du français, expriment leur engament dans un court texte argumentatif sous forme de lettre officielle.

2 Corpus et méthodologie

Pour répondre à notre problématique, nous avons constitué un corpus de productions écrites d'apprenants sinophones du français composé de 10 textes. Agés de 20 ans en moyenne, les apprenants suivent des cours de FLE au Centre Universitaire d'Etudes Françaises (CUEF) à Grenoble et ont atteint le niveau B2 en français. Ils sont en France depuis environ un an. Leurs productions ont été contrastées à celles des groupes de contrôle composés de scripteurs sinophones et francophones natifs. Les sinophones natifs (10 participants) sont étudiants l'Université de Suzhou (Chine) dans le domaine des sciences humaines. Quant aux francophones (10 participants), ils sont étudiants à l'Université Grenoble Alpes en 3e année de Licence Sciences du langage. D'après notre questionnaire, ils ne pratiquent pas d'autres langues que leur L1 dans la vie de tous les jours et ont le même âge que les apprenants.

Les productions ont été réalisées en situation d'examen dans un temps limité à une heure. La consigne donnée en français correspond à :

> *Vous habitez dans une ville qui organise chaque année un grand concert gratuit pour marquer la fin de l'été. Pour des raisons financières, votre ville annonce qu'elle veut supprimer cet événement musical. Vous écrivez au maire de la ville pour le persuader, à l'aide d'arguments et d'exemples précis, des avantages culturels et touristiques que ce concert représente. Vous insistez également sur l'intérêt économique de cette manifestation pour les commerçants et les artistes de la région. (250 mots)*

Cette consigne a été empruntée aux épreuves du DELF pour le niveau B2. Elle a été traduite en chinois pour les scripteurs sinophones. Le nombre de mots a été limité à 250 pour les textes rédigés en français et à 500 caractères pour le chinois. Notre corpus est donc bilingue et comparable. Il représente 11914 mots/caractères pour le chinois.

Type de scripteurs	**Nombre de textes**	**Nombre de mots/caractères**
Francophones natifs	10	2875
Sinophones natifs	10	5914
Apprenants	10	3125
TOTAL	**30**	**11914**

TAB. 1 : Taille et nature du corpus

Les productions manuscrites ont été transformées au format Word et XML. Les productions en chinois ont été glosées afin d'avoir une information sur la nature grammaticale de l'élément et sa traduction en français.

3 Résultats

Dans l'ensemble des données, on distingue trois formes qui permettent aux scripteurs de montrer leur implication dans le texte. Il s'agit des pronoms personnels, des adjectifs possessifs et de la forme impérative du verbe.

3.1 Groupes de contrôle

Dans les productions des groupes de contrôle, les pronoms personnels occupent la place majoritaire chez les francophones (63%) et les sinophones (96%). Les adjectifs possessifs représentent 33% chez les francophones et 4% chez les sinophones, et les formes impératives sont présentes uniquement dans les productions des francophones à hauteur de 4% (figure 1).

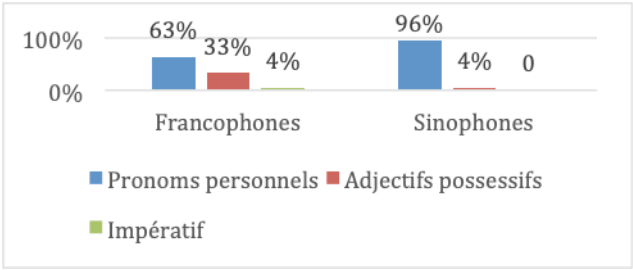


FIG. 1 : Moyens d'implication du scripteur utilisés par les groupes de contrôle

Les pronoms personnels utilisés sont équivalents en français et en chinois avec une préférence pour le pronom « vous » chez les francophones. En revanche, dans les textes en chinois le pronom le plus fréquent est 我 équivalent de « je ». Ce pronom occupe la deuxième place chez les francophones natifs. Les autres pronoms comme « nous » et « on », sont moins fréquents dans les productions des deux groupes. Quant aux adjectifs possessifs, ils sont moins fréquents dans les productions en chinois que dans celles produites par les francophones. Cette différence résulte du statut du déterminant dans les deux langues.

3.2 Apprenants

Quant aux apprenants, les moyens d'implication du scripteur se limitent aux pronoms personnels et aux adjectifs possessifs avec une dominance des premiers (figure 2).

Pour ce qui est du répertoire des pronoms, le pronom « je » est le plus fréquent. Il est suivi par « nous », « vous » et « on ». En ce qui concerne les adjectifs possessifs, ils sont employés aussi systématiquement que chez les francophones monolingues.

4 Discussion

Nous constatons donc une influence de la L1 des apprenants sur le choix du pronom personnel qui sert à montrer son engagement dans le texte. Les apprenants se focalisent davantage sur le scripteur que sur le destinataire ne souhaitant probablement pas être trop invasifs. Les formes

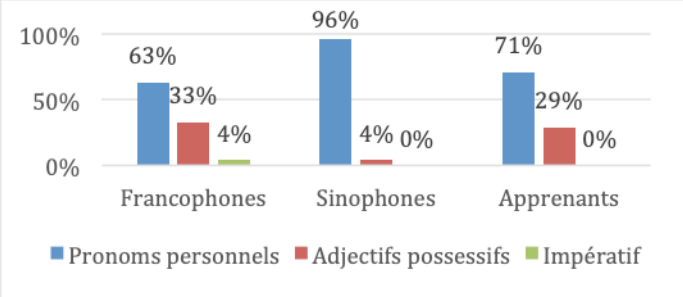


FIG. 2 : Moyens d'implication du scripteur utilisés par l'ensemble de scripteurs

impératives ne sont pas employées. En revanche, les adjectifs possessifs sont utilisés aussi systématiquement que par les francophones. Ce résultat est probablement dû au statut obligatoire de détermination en français. L'analyse d'autres critères rhétoriques est en cours et nous souhaiterions approfondir nos analyses pour dégager des spécificités rédactionnelles de ce public.

Références bibliographiques

Bi, X. (2016). *Rhétorique de la dissertation : étude contrastive des conventions d'écriture académique en français et en chinois* (Thèse de doctorat). Université Sorbonne Paris. Repéré à `http://tel.archives-ouvertes.fr`

Connor, U. (1996). *Contrastive rhetoric : cross-cultural aspects of second-language writing.* Cambridge, Royaume-Uni de Grande-Bretagne et d'Irlande du Nord, États-Unis d'Amérique : Cambridge university press.

Hidden, M.-O. (2013). *Pratiques d'écriture : apprendre à rédiger en langue étrangère.* Paris : Hachette français langue étrangère.

Wang, W. & Wen, Q. (2002). L1 use in the L2 composing process : an exploratory study of 16 Chinese EFL writers. *Journal of Second Language Writing, 11*, 3.

Constitution d'un corpus d'apprenants du FLE –enjeux et pistes de recherche

Magdalena Augustyn [1], Thi Thu Hoai Tran [2]et Rui Yan [1].
[1]LIDILEM, Université Grenoble Alpes
[2]GRAMMATICA - Université d'Artois
Magdalena.Augustyn@univ-grenoble-alpes.fr, tthoai.tran@univ-artois.fr, Rui.Yan@univ-grenoble-alpes.fr

Mots clés : corpus d'apprenants, FLE, FOU, écrits académiques, analyse d'erreurs, linguistique de corpus

Résumé : Nous nous proposons de présenter le projet de constitution d'un corpus des écrits académiques des apprenants allophones de niveau B1 à C1 du CECRL. Il s'agit d'un recueil de productions diversifié aussi bien sous l'angle du profil d'apprenant (un corpus d'apprenants multi-L1, à différents niveaux de compétence), que des genres (compte-rendu de lecture, rapport, dissertation, travail d'analyse, résumé, etc.) et des supports (dans un premier temps : écrit manuscrit, écrit tapuscrit). Une réflexion sur des pistes de didactisation dans une perspective du Français sur Objectif Universitaire est parallèlement menée.

1 Enjeux et objectifs

Nous nous proposons de présenter le projet de constitution d'un corpus des écrits académiques des apprenants allophones du FLE, de niveau B1 à C1 du CECRL, ainsi qu'une première étude basée sur les données recueillies.

Alors qu'en anglais langue étrangère il existe de nombreuses ressources et études basés sur les corpus d'apprenants (De Cock et Tyne, 2014, Granger *et al.*, 2015), l'intérêt accordé à ce type de corpus et leur prise en compte dans l'enseignement/apprentissage du FLE sont plus récents[1]. Nous souhaitons pour notre part contribuer à la recherche dans ce domaine en proposant une mise en place d'un corpus écrit d'apprenants du FLE, ainsi que des pistes de didactisation dans une perspective du Français sur Objectif Universitaire.

Notre premier objectif est de constituer un large corpus de productions d'apprenants du FLE et de le mettre à disposition de la communauté des chercheurs et enseignants. Il s'agit d'un recueil de productions diversifié aussi bien sous l'angle du profil d'apprenant (un corpus d'apprenants multi-L1, à différents niveaux de compétence), que des genres (compte-rendu de lecture, rapport, dissertation, travail d'analyse, résumé, etc.) et des supports (dans un premier temps : écrit manuscrit, écrit tapuscrit). Cette première étape comprend également une élaboration d'un protocole de recueil et d'annotation du corpus afin de permettre un enrichissement progressif de la base avec des données présentant les mêmes propriétés que le corpus de base et permettant ainsi

1. Par exemple, des corpus de productions orales des apprenants du FLE : dans le cadre du projet IPFC (Interphonologie du français contemporain : `https://www.projet-pfc.net/`) ou du projet InterFra (Interlangue française –développement, interaction and variation : `https://spraakbanken.gu.se/eng/resource/interfra`).

par la suite une exploitation et analyses comparatives fiables. Cette ressource sera expérimentée auprès d'étudiants qui souhaitent s'intégrer dans les universités françaises dans l'objectif de les aider à se familiariser aux spécificités des écrits académiques.

2 Constitution du corpus : état actuel

A ce jour, le corpus au format xml/TEI est composé de 93 travaux universitaires (environ 130000 mots) d'étudiants allophones de Master 1, Master 2 de l'Université Grenoble Alpes en sciences humaines, du DU Lettres, Langue Française, Communication, formation dispensée par le CUEF de l'Université Grenoble Alpes, ainsi que du DU FLEPES (Diplôme Universitaire pour la Préparation des Etudes Supérieures) à l'Université d'Artois.

Le corpus comporte des métadonnées avec des éléments socio-biographiques des apprenants (David & Aleksandrova, 2017) permettant par la suite d'effectuer des recherches dans le corpus selon divers critères (âge, langue(s) maternelle(s), genre textuel, etc.). Afin de garantir des possibilités de recherche plus avancées, un travail supplémentaire à la constitution du corpus est effectué en parallèle : une annotation de la structure des textes, ainsi que le repérage des segments correspondant aux éléments susceptibles de biaiser les requêtes portant sur l'interlangue comme les citations, les traductions littérales ou les passages et exemples en langue étrangère. Le corpus sera à terme mis à disposition en accès libre, cet enrichissement du corpus constituant ainsi une plus-value qui en fera une source d'information linguistique pour des recherches et développements ultérieurs (Garside, Leech, McEnery, 2013).

3 Pistes d'exploitation

Sur le plan didactique, le corpus d'apprenants nous permettra tout d'abord de relever les erreurs des apprenants. Ces dernières sont en cours de repérage d'après une typologie d'erreurs préalablement établie (Yan, 2017) et feront l'objet d'un étiquetage formel. De manière générale, trois classes d'erreurs peuvent être distinguées : au niveau morphologique, au niveau de la grammaire et de la syntaxe, au niveau sémantique et des cooccurrences lexicales. Au sein de chaque classe, différents types d'erreurs peuvent être repérés, par exemple, pour la dernière classe, on trouve des erreurs de sens, de collocation et de cooccurrence.

Dans le cadre d'une première étude basée sur corpus, nous portons une attention particulière aux constructions verbales impersonnelles, de type : *il convient d'étudier, il est à noter que, il en résulte que*. Il s'agit des expressions semi-figées particulièrement présentes dans les écrits académiques et servant à véhiculer les différentes stratégies discursives propres à ce genre d'écrit. Par l'intégration de ces formes discursives, nous souhaitons sensibiliser les apprenants aux spécificités rhétoriques dans les écrits scientifiques.

Un corpus d'apprenants comme le nôtre qui regroupe les écrits des apprenants de différentes langues maternelles nous permettra également d'effectuer des comparaisons, quantitative et qualitative, entre les productions de différents groupes de locuteurs non-natifs (cf. la méthode *Contrastive Interlanguage Analysis* (CIA), Granger, 2002). Il est à notre sens important d'appliquer ces

deux méthodes d'exploitation dans les recherches sur les corpus d'apprenants afin de mieux comprendre les difficultés des apprenants et d'expliquer certains phénomènes observés en distinguant notamment des facteurs spécifiques aux L1 et des facteurs plus universaux. Ce type de recherche permet par la suite de mieux orienter les pratiques d'enseignement. En effet, l'exploration des différents aspects de l'acquisition d'une L2 sur corpus donnera lieu à terme à des applications pédagogiques ciblées dans le cadre de la didactique du Français sur Objectif Universitaire plus particulièrement, tels que l'apprentissage sur corpus ou l'élaboration d'outils d'aide à la rédaction.

Références bibliographiques

David, C., & Aleksandrova, T. (2017). Acquisition de la compétence de production écrite en FLE par des apprenants chinois : l'exemple de l'essai argumenté. *Actes des 9èmes Journées Internationales de la Linguistique de corpus.*

De Cock, S., Tyne, H. (2014). Corpus d'apprenants et acquisition des langues. *Recherches en Didactique des Langues et Cultures* : les Cahiers de l'acedle, L'association des chercheurs et enseignants en didactique des langues étrangères, Recherches en Didactique des Langues et des Cultures, 11(1), pp.137-168.

Garside, R., Leech, G., and McEnery, T. (2013). *Corpus Annotation : Linguistic Information from Computer Text Corpora*, 3rd edn. Abingdon, UK : Routledge.

Granger, S. (2002). A Bird's-eye View of Computer Learner Corpus Research. In S. Granger, J. Hung, & S. Petch-Tyson (Éd.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Benjamins : Amsterdam & Philadelphia, 3-33.

Granger, S., Gilquin, G., Meunier, F. (Eds.) (2015). *The Cambridge Handbook of Learner Corpus Research*, Cambridge University Press.

Yan, Rui. (2017). *Étude des constructions verbales scientifiques dans une perspective didactique : utilisation des corpus dans le diagnostic des besoins langagiers du FLE à l'aide des techniques de TAL* (Thèse de doctorat en sciences du langage). Université Grenoble Alpes, Grenoble.

Constitution et exploitation d'un corpus d'arabe tunisien

Fatma Ben Barka Messaoudi .
LLL (UMR 7270), Université d'Orléans
Fatma.messaoudi@univ-orleans.fr

1 Introduction

Dans cette étude, notre regard sera focalisé sur l'ensemble des procédures, allant du recueil de données jusqu'à la phase de transcription, qui a été mis en œuvre afin de constituer un corpus échantillonné et diversifié de 20 heures d'enregistrements de l'arabe tunisien effectués à Orléans et en Tunisie, dans le cadre de notre étude doctorale (en cours) sur le subjonctif.

2 Constitution du corpus

2.1 Données orales en arabe tunisien : des corpus fantômes[1] ?

Depuis quelques années, les chercheurs ont commencé à s'intéresser aux langues peu dotées dans l'objectif de les documenter, de les décrire et de les analyser selon différentes approches. Ayant conscience de l'accroissement de la linguistique de corpus, des acteurs de la recherche sur l'AT se sont tournés principalement vers les médias (Hamdi & al. 2013) et vers les réseaux sociaux (Zamiti 2015, McNeil & Faiza 2011, Younes & al. 2014 ; 2015) afin de construire des corpus en AT. Nonobstant, ces données ne sont ni disponibles, ni accessibles à l'ensemble de la communauté scientifique. Nous pouvons dire ainsi que ce sont des « corpus totalement fantômes qui fondent certains travaux sans qu'aucune information ne précise les raisons de l'absence de l'accès aux données, pourtant seule garantie d'un travail scientifique en principe ouvert à la falsification. » (Abouda & Baude 2006) De part ce fait, il a fallu mener nous-mêmes le travail de collecte des données.

2.2 Démarche de la constitution du corpus

Pour recueillir un échantillon authentique d'arabe tunisien, nous nous sommes inspirée des démarches adoptées par Moukrim (2010) et Ben Ahmed (2017) et reposée fondamentalement sur les caractéristiques quantitatives et qualitatives de notre corpus ESLO, et plus spécifiquement ESLO2 (cf. tableau 1).

S'agissant d'une recherche contrastive (français, arabe tunisien) mixte (à la fois micro-diachronique et synchronique), les choix présidant à la construction du corpus d'AT ont été guidés par un souci de comparabilité avec notre sous-corpus d'étude ESLO.

Pour le mode de recueil des données, nous avons favorisé l'entretien en face-à-face, « situation certes très formelle, mais qui avait l'avantage d'être (...) contrôlable » (Abouda & Baude 2009 : 134). Néanmoins, dans l'objectif d'assurer un certain équilibre au sein de notre corpus, nous

1. Reprise de la notion de corpus fantômes avancée par ABOUDA L., BAUDE O. (2006).

	Genre interactionnel	ESLO1 & ESLO1-MD	ESLO2 & ESLO2-MD
Durée (en min)	Entretiens	2234	2230
	Repas	224	236
	Conférences	215	211
	Total	2673	2677
	Total général	5350 (89h & 10 min)	
Nombre de locuteurs	Entretiens	35	43

TAB. 1 : composition du sous-corpus ESLO

avons intégré, à hauteur de 20%, deux autres genres interactionnels de « contrôle », i.e. les repas et les cours universitaires.

Quant au questionnaire, nous nous sommes appuyée sur les principaux thèmes retenus par ESLO (logement, travail, loisirs, langues, Orléans), tout en abordant d'autres sujets (par exemple celui de la révolution tunisienne) susceptibles de faire parler les locuteurs tunisiens, en visant les contextes propices à l'apparition des formes verbales subjonctives.

Suivant la démarche d'ESLO, indispensable pour rendre le corpus disponible et interopérable, nous avons effectuée une documentation détaillée de toutes les informations à propos les enregistrements, leurs contextes de production et les témoins.

Nous avons essayé, autant que possible, de respecter l'unité de lieu. Nous avons donc débuté notre enquête à Orléans. Mais, afin de répondre aux contraintes sociologiques, nous étions obligée d'élargir notre terrain d'enquête en réalisant quelques enregistrements en Tunisie.

En somme, notre corpus global est composé de deux principales parties : 7h d'enregistrements sélectionnés du corpus de Ben Ahmed (2014) et 13h enregistrements collectés par nous-mêmes (2017), un volume jugé suffisant pour mener notre étude contrastive sur le subjonctif.

	AT 1 (Ben Ahmed : 2014)	AT 2 (Ben Barka : 2017)
Lieux d'enquête	Orléans –Tunisie	
Nombre d'heures	7	13
Nombre de mots	56028	108705
Nombre de locuteurs	11	17
Situations de communications	entretien face à face –repas –cours universitaires	

TAB. 2 : le corpus en question

2.3 Transcription

Une fois le corpus constitué et afin de faciliter son exploitation, un travail préalable de transcription a été nécessaire. Cette phase indispensable dans le processus de traitement du corpus a été effectuée en deux étapes : une transcription brute (version A) des enregistrements d'AT 2 et une relecture (version B) des enregistrements d'AT 1.

La première étape de transcription a soulevé plusieurs interrogations : sur quel outil transcrire ? Quels système et conventions de notation adopter ? Quel mode de transcription choisir ?

Le manque d'outils dédiés à la transcription de l'AT nous a poussée à annoter nos données brutes sous TRANSCRIBER, un logiciel d'aide à l'annotation des données audio permettant, grâce à l'encodage UTF8, de transcrire plusieurs langues (européennes et/ou non européennes). Cet outil, qui a été développé par Claude Barras et Edouard Geoffroy de la Direction Générale de l'Armement (DGA), facilite la visualisation, la manipulation et le traitement des sources sonores. Ensuite, en tenant compte de la spécificité de l'AT en tant qu'une langue à tradition orale, qui se distingue de l'arabe classique phonologiquement, morpho-syntaxiquement et lexicalement, nous avons favorisé la graphie latine au détriment de la graphie arabe. Nous avons opté pour ce choix afin d'écarter les contraintes de l'écriture arabe (de droite à gauche) et de constituer un corpus de référence de l'arabe tunisien partageable et lisible par les non-natifs. Puis, pour le choix des conventions de transcription, en l'absence d'une tradition orthographique de l'AT, nous nous sommes reposée fondamentalement sur les recommandations de l'INALCO (1996 –1998). Quand il s'agissait des phénomènes de l'oralité, nous avons adopté les conventions proposées par le LLL dans le cadre du projet ESLO. Enfin, afin de réaliser une transcription alignée avec le son, nous avons sélectionné le mode orthographique (usuel) d'inspiration phonologique où les dimensions morphosyntaxiques des énoncés sont conservées, ce qui offre une lecture rapide et une simplification de l'exploitation du corpus par les logiciels de concordance.

Lors de la deuxième étape de transcription, il a été question de relire et corriger la version A de Ben Ahmed (2014). Nous étions donc emmenée à revoir la segmentation des tours de parole, vérifier l'orthographe et les conventions utilisées et revenir sur la transcription des phénomènes de l'oralité (bruits, silences, hésitations). Lors de cette phase d'annotation, les problèmes que nous avons rencontrés ont été majoritairement liés aux variations régionales. L'impact de l'empreinte des transcriptrices (YBA et FBB) s'est manifesté par :

(1) des variations de perception
YBA : nažam
FBB : ennežžem (fr : je peux)

(2) des variations de segmentation
YBA : b-dhabt
FBB : bi-el-dhabt (fr : exactement)

(3) des variations graphiques
YBA : ena w xouya
FBB : ena ou xouya (fr : moi et mon frère)

Ces exemples mettent en avant les difficultés à représenter graphiquement l'arabe tunisien, en l'absence d'une orthographe standard et d'une description grammaticale de cette langue.

3 Conclusion

Dans ce papier, nous avons exposé les choix méthodologiques et techniques que nous avons opérés lors de la construction et la transcription d'un corpus échantillonné et diversifié de 20 heures d'enregistrements du parler tunisien afin de répondre aux contraintes rencontrées et de constituer un corpus de référence de l'arabe tunisien accessible à tout chercheur intéressé et pouvant faire l'objet des prochaines études sur cette langue sous documentée.

Références bibliographiques

Abouda L., Baude O. (2006). Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des ESLO, CORAL –Université d'Orléans.

Abouda, L., Baude, O. (2009). Du Français Fondamental aux ESLO, Bruxelles, Mondada, Simon, Traverso Grand corpus de français parlé, Bilan historique et perspectives de recherche, *Cahiers de Linguistique Revue de sociolinguistique et de sociologie de la langue française 33/2*, EME, Louvain, 131-146.

Ben Ahmed, Y. (2017). Constitution d'un corpus d'arabe tunisien parlé à Orléans, Actes de l'atelier « Diversité Linguistique et TAL » (DiLiTAL 2017), 62-69.

Hamdi, A., Boujelbane, R., Habash, N., & Nasr, A. (2013). Un système de traduction de verbes entre arabe standard et arabe dialectal par analyse morphologique pro fonde. Traitement Automatique des Langues Naturelles. 396-406.

McNeil, K., Faiza, M. (2011). Tunisian Arabic Corpus : Creating a written corpus of an "unwritten" language, *Proceedings of Workshop on Arabic Corpus Linguistics*, Lancaster, UK.

Moukrim, S. (2010). Sur la constitution de Corpus de deux langues à tradition orale (l'arabe marocain et le berbère tamazight), parlées à Orléans, Azzopardi, S. (éd), Corpus, données, modèles. Cahiers de Praxématique, 54-55 PULM. Montpellier, 367-378.

Younes, J., Souissi, E. (2014). A quantitative view of Tunisian dialect electronic writing, *Proceedings of the 5th International Conference on Arabic Language Processing*, Oujda, Morocco, 63-72.

Younes, J., Achour, H. & Souissi, E. (2015). Constructing Linguistic Resources for the Tunisian Dialect Using Textual User-Generated Contents on the Social Web, *Proceedings of the 1st International Workshop on Natural Language Processing for Informal Text (NLPIT 2015) In conjunction with The International Conference on Web Engineering* (ICWE 2015), Rotterdam, The Netherlands.

Zamiti, A. (2015). Analyse diachronique de concepts politiques dans un corpus en tunisien issu de Facebook. Mémoire de fin d'études dirigé par Mathieu Valette. Université Paris III Sorbonne Nouvelle.

La technique de stylométrie à la base de l'analyse informatique du rythme du texte

Elena Boytchuk et Olga Belyaeva .
Université Pédagogique d'Etat Oushinsky
elena-boychouk@rambler.ru, olbelyaeva@yandex.ru

1 Introduction

Le rythme de la prose ayant à sa base des répétitions de différents types représente un phénomène complexe et se réalise aux différents niveaux de langue. Un grand nombre d'études théoriques existantes dans le domaine du rythme, nécessitant une approche statistique, l'utilisation de méthodes d'analyse quantitative, le traitement de texte automatisé, n'entraînent pas la mise en œuvre de ce type d'analyse. La technique de stylométrie n'est appliquée que dans le processus du traitement manuel du texte [1]. L'analyse du rythme lui-aussi est effectuée dans la plupart des cas manuellement, ce qui réduit l'efficacité du traitement de texte et ne permet pas d'obtenir des résultats précis. Le problème du traitement automatique du rythme du texte est pertinent du fait que de nombreuses études théoriques sont consacrées à l'étude du rythme du point de vue de divers aspects de sa manifestation, mais que les méthodes et moyens de son traitement ne sont pas suffisants pour effectuer une analyse rythmique automatisée complexe. Dans le même temps, la méthode principale de traitement des textes en langues naturelles est une méthode quantitative effectuée manuellement, ce qui réduit l'efficacité de la recherche.

2 Corpus et méthodologie

2.1 Corpus

Dans le cadre du traitement automatique du texte il existe beaucoup d'études consacrées à la réalisation de l'aspect phonétique, lexique et grammatical.

Les aspects principaux de l'analyse automatique du texte sont décrits dans la recherche de F.J. Sanchez Perez [2]. L'auteur porte l'accent sur la méthodologie, prenant comme méthodes principales la détection et le traitement quantitatif des données, l'étude des données lexicométriques (classement des mots selon les paramètres différents –hiérarchiques, quantitatifs, alphabétiques, etc.). L'analyse de la structure rythmique du texte selon J.-Ph. Massonie [3] est basée sur la répétition des mots. L'utilisation de cette option rend possible la division du texte en parties selon le contenu. Les chercheurs E. Greene, T. Bodrumlu, K. Knight [4] effectuent l'analyse du rythme du texte poétique à la base de l'analyse de la mesure.

Ayant comme base les études de R. Ghiglione, les chercheurs P. Molette et A. Landré [5] ont créé la méthode d'analyse du texte « Tropes ». Ce programme permet de déterminer le champ sémantique du texte. Il effectue l'analyse de la structure morphologique et différencie les types de discours (narratif, descriptif, explicatif, argumentatif).

Les approches présentées ci-dessus accomplissent l'analyse de différents moyens de rythme. Dans le cadre de la conception d'analyse du rythme (au niveau phonétique, lexique et grammatical) adoptée dans cet article, la dernière méthode se présente comme la plus efficace. Tout de même elle n'est pas suffisante pour déterminer le degré de la rythmisation du texte ce qui s'explique par la complexité du rythme réalisée aux différents niveaux de langue.

2.2 Méthodologie

L'approche informatique "Rythmanalyse" créée pour travailler avec le texte français permet d'apprécier le rythme du texte prosaïque non seulement du point de vue de la grammaire mais aussi des marques phonostylistiques (assonance, allitération, rime et unités rythmiques), lexiques et grammaticales (répétitions de types différents –épanalepse, redoublement, anadiplose et autres.). Pour créer cet instrument, on a utilisé Qt Framework et la technologie Qt Quick. Il peut être exécuté sur diverses plates-formes, y compris MS Windows, GNU / Linux et Mac OS X.

Marques phonostylistiques

Les conditions du traitement du texte sont déterminées par les règles de prononciation et de division en syllabes dans la langue française.

A. Division en unités rythmiques
   La division du texte en syllabes est basée sur les règles de lecture : la combinaison des sons et des diphtongues (classifiés par le programme comme des syllabes : iè, ieu, eai ··· etc.) et les différentes positions des consonnes (dont la prononciation dépend des sons qui l'entourent).
   Sur cette base des règles de la division du texte en syllabes aussi bien que sur la base des règles de la ponctuation et de l'emploi des conjonctions de coordination et de subordination, est créé l'algorithme de la détermination des unités rythmiques du texte.
   Bien sûr cette division n'est pas exacte, mais elle facilite beaucoup la procédure de traitement du texte et donne la possibilité de déterminer la quantité de syllabes qui à son tour permet de révéler les unités comprenant une quantité de syllabes identique (ce qui est étroitement lié au rythme du texte : les unités équivalentes du point de vue de leur composition syllabique ou les unités représentants une succession de syllabes avec une différence d'une ou deux syllabes sont classées comme les plus rythmiques). La couleur permet d'évaluer visuellement le texte et d'y trouver facilement les unités ayant le même nombre de syllabes ou un nombre approchant (c'est à dire une différence de deux ou trois syllabes).

B. Assonance et allitération
   Dans l'objectif de déterminer les consonnes et les voyelles du fragment de texte proposé, des règles permettant de différencier les consonnes et les voyelles prononcées et non prononcées en fonction de leur position et de leurs combinaisons ont été formulées. Afin de faciliter le travail sur de longs extraits, cet instrument permet de visualiser la fréquence d'une consonne en la surlignant dans le texte étudié.

On révèle des cas d'assonance comprenant parfois un groupe de sons qui ont des ressemblances du point de vue phonétique, par exemple, le signe [a] représente le groupe de sons suivants [a], [ã], [wa], [ua], [ya], [aj], [ja], le signe [e] représente les sons [ej], [je], [ɛ], [jɛ], [ɛj], [yɛ], [uɛ] etc.

C. Rime
L'analyse de la rime s'effectue de point de vue de la division en trois types : rime pauvre (reprise de la même voyelle accentuée : canaux - vaisseaux), rime suffisante (reprise d'un groupe constitué d'une voyelle et d'une consonne : final / amical) et rime riche (reprise d'au moins trois phonèmes : cheval / rival). La requête de la rime est accompagnée de l'indication de la quantité des mots rimés dans le texte. Afin de faciliter le travail sur de longs extraits, l'instrument permet de visualiser la fréquence des mots rimés en les surlignant dans le texte étudié.

Aspect lexical et grammatical

Au niveau lexical cet instrument permet de dégager "les mots-clés" du texte ou de son fragment en indiquant la fréquence de leur emploi. L'approche ne réalise que l'analyse quantitative des répétitions des parties du discours indépendantes essentielles (nom, verbe, adjectif, adverbe).

L'analyse de l'aspect grammatical est basée sur une identification de l'anaphore, épiphore, symploque, des termes homogènes, sur la fréquence des propositions ayant différents buts communicatifs (question, exclamation, réticence), anadiplose, réduplication, épanalepse, polysyndète.

Cet instrument permet une détermination automatique des répétitions à tous les niveaux de langue : phonétique, lexical, stylistique et grammatical [6].

Pour faciliter le travail avec les textes en anglais nous avons créé un instrument permettant chercher dans le texte et déterminer la quantité des procédés stylistiques suivants : anaphore, épiphore, symploque, anadiplose, épanalepse, réduplication.

Le linguiste peut recevoir des informations sur les aspects lexicaux utilisés dans le texte, élaborer des théories sur l'affiliation du texte d'un auteur concret, évaluer la qualité de la traduction et analyser le rythme du texte. L'instrument est une page Web qui a été implémentée sur le langage JavaScript en utilisant HTML et CSS.

3 Résultats

Une étude expérimentale de l'efficacité de ces applications a été menée à la base des travaux des auteurs français et anglais. Pour cette étude, 100 œuvres des auteurs suivants ont été sélectionnées : G. Flaubert, G. de Maupassant, Stendhal, E. Zola, A. Maurois, O. de Balzac, W. Scott, J. Austen, Ch. Dickens, E. Brontë, G. Eliot. L'utilisation de l'application « Rythmanalyse »

et d'un outil d'analyse des procédés rythmiques dans les textes anglais a permis d'atteindre les résultats suivants :

- Le temps consacré au traitement du texte à l'aide des méthodes proposées a été considérablement réduit. L'analyse manuelle des travaux a pris 15 fois plus de temps que l'utilisation des méthodes informatisées d'analyse du rythme.
- Les outils peuvent réduire l'impact du facteur humain sur les résultats de l'étude. Les cas d'erreurs et les erreurs commises par le chercheur lors du traitement manuel du texte sont exclus.
- Le traitement du texte effectué en termes de structure rythmique a permis d'identifier les spécificités du style de chaque auteur et de déterminer les critères permettant de le reconnaître.
- L'étude a montré que les œuvres de G. de Maupassant parmi les textes des auteurs français étudiés (de Stendhal, O.de Balzac, G. Flaubert, E. Zola, A. Maurois) sont les plus rythmées. Les procédés les plus fréquemment utilisés sont la rime et l'assonance au niveau phonostylistique, l'épanalepse, l'anadiplose et le redoublement au niveau grammatical.
- Parmi les textes anglais les oeuvres d'E. Brontë sont les plus rythmées surtout du point de vue de la fréquence des moyens rythmiques au niveau grammatical. Les moyens les plus fréquemment utilisés par l'auteur sont l'anaphore, l'épanalepse et les termes homogènes.

Références bibliographiques

[1] Zenkov Andrei V. (2017). A Method of Text Attribution Based on the Statistics of Numerals, Journal of Quantitative Linguistics, [version en ligne]

[2] Sanchez Perez F. J. (1993). Qu'est-ce que l'analyse relationnelle informatique des textes ? *La Revue Informatique et Statistique dans les Sciences humaines* XXIX, 1 à 4. Paris.

[3] Massonie, J.-Ph. (1990). Analyse informatisée des textes. *Annales littéraires de l'Université de Besançon.* 154 p.

[4] Greene E., Bodrumlu T., Knight K. (2010). Automatic Analysis of Rhythmic Poetry with Applications to Generation and Translation. EMNLP.

[5] Molette P. (2009). De l'APD à Tropes : comment un outil d'analyse de contenu peut évoluer en logiciel de classification sémantique generalist. *Psychologie Sociale et Communication*.

[6] E. Boychuk, I. Paramonov, N. Kozhemyakin, and N. Kasatkina (2014). Automated approach for rhythm analysis of French literary texts. *Proceedings of 15th Conference of Open Innovations Association FRUCT.IEEE,* pp. 15–23.

REMERCIEMENTS

L'étude a été réalisée avec le soutien financier de la Fondation russe pour la recherche fondamentale (RFBR) dans le cadre du projet de recherche №19-07-00243.

EFL writing skills brought from high school: Corpus-based research on English majors' take-home essays

Viola Kremzer .
University of Pécs
Kremzer.viola@pte.hu

1 Introduction

High school students worldwide applying for a BA in English Studies program are required to reach upper-intermediate English as a foreign language (EFL) proficiency, well-developed language skills, and a broad mental lexicon with a wide variety of words and word clusters. According to the Common European Framework of Reference for Languages (CEFR), upper-intermediate level is described as level B2 or B2+. By being admitted to the English Studies program, students become members of the academic community; and they are required to be familiar with academic genres, including essays, reviews, articles, and letters, and they need to have knowledge and skills to create such texts and structure them accordingly. Later, they can apply their obtained knowledge in various academic genres in the university context.

In Hungary, application for any BA in English Studies Program requires students to accomplish the advanced school-leaving examination or any B2 level (or above) language examination certificate accredited by the Government. Obtaining a score 60% on the exam equals a B2 CEFR level. The present study is embedded in research on academic discourse (Csomay, 2005; Biber & Gray, 2010), more specifically essay writing (Bailey, 2011; Creme and Lea, 2008) and aims at researching first-year English majors' writing skills at the beginning of their studies in tertiary education. Entering higher education requires students to socialize into the academic community (Duff, 2010; Swales, 1993), where they acquire the ability to speak and write to the academic audience by getting to know more about academic language use, style, and structure. However, how much do they need to learn and improve?

2 Corpus and methodology

The participants were 92 first-year students attending three 'Reading and Writing Skills' courses, which serve as introductory courses to academic skills focusing on reading and writing as well as language development in general. The courses were tutored by three professors experienced in skill development at the Institute of English Studies. To examine first-year English majors' writing skills in the Hungarian context, I conducted a study guided by the following questions:

RQ1 How can the vocabulary of the take-home essays written by first-year English majors be described in a corpus-based approach?
RQ2 What do syntactic measures indicate?
RQ3 What characterizes the cohesion and coherence of take-home essays?
RQ4 What characterizes the narrativity of student essays?
RQ5 What readability level do the essays reach?

2.1 Corpus

In the present study, the corpus consisted of 92 texts, and it covered 36,311 tokens with the shortest essay being 258- and the longest one being 606-words (mean=394.68, SD=68.32). The topic of the essays was being an English major at the University of Pécs and data collection took place in the third week of the fall semester in 2017. According to the prompts given, essays were required to consist of four paragraphs, each with specific focus: a) story of the decision to become an English major, b) expectations as an English major, c) importance of this knowledge in the future, and d) an inspiring story about the impact of English on a person's life. Students were asked to write 300-400 words without the inclusion of an introduction or a conclusion. The task was given as home assignment with clear instructions on the structure and detailed evaluation criteria. In the study, the essay corpus compiled from these pieces of work was representative of academic essays and compiled for analyzing students' writing skills (Flowerdew, 2012). After deleting personal information (Dörnyei, 2007) from the essays for ethical considerations, the corpus was assembled following pre-determined coding procedures.

2.2 Methodology

After establishing the measures of analysis, each essay was entered into VocabProfile VP-Compleat software program on Tom Cobb's Compleat Lexical Tutor website to create individual text profiles as well as a joint corpus of the 92 student texts was created and analyzed for an overall lexical profile of the student texts. Data analysis included lexical richness, lexical density (LD), core general English vocabulary as represented by the New General Service List (NGSL; Browne, Culligan, Phillips, 2013), academic vocabulary on the basis of the New Academic Word List (NAWL; Browne, Culligan, & Phillips, 2013), and words no in any of these lists categorized as off-list.

Vocabulary analysis was followed by the examination of the 92 essays in the Coh-Metrix Test Easability Assessor, a text-analysis software available online calculating linguistic features of a text shorter than 1,000 words. I analyzed the corpus for six factors: narrativity, syntactic simplicity, word concreteness, referential cohesion, deep cohesion, and the Flesch-Kincaid Grade Level (FKGL). These scores provide a profile of an essay that mirrors how easy it is to read and what aspects ought to be improved in the writing. I used the IBM SPSS platform to analyze data obtained about vocabulary and readability profiles of the student take-home essays. The focus of SPSS analysis was to determine statistical differences, correlations and occurrent outliers.

3 Results

At this stage of the students' vocabulary development, the academic expectation is a strong upper-intermediate level, but the lexical analysis of students' essays shows a different situation. The lexical frequency profiles of the essays produced by the 92 first-year students ranged between 76.26% and 93.35% (M=87.43%, SD=3.01), the type-token ratios between 0.39 and 0.56 (M=0.48, SD=0.04) and the lexical density figures between 0.42 and 0.54 (M=0.47, SD=0.03). These numbers demonstrate considerable individual variability among home assignment essays.

Variation can be explained by diversity in participants' lexical knowledge. The academic vocabulary profiles range between 0% and 2.71% (M=0.75, SD=0.51). The reason for academic words observed to be less frequent may be students' limited language proficiency level or their lack of academic lexical knowledge since they were in their first semester in the academic context. Sophisticated vocabulary use is a fundamental element of academic writing (Luey, 2006); however, slang and colloquial expressions also occured in the corpus.

Findings show a high rate of narrativity (M=87.52%, SD=10.82), indicating that the texts were story-like and described personal anecdotes. It shows that the task was accomplished by the participants as required by the prompts. Referential cohesion ranged from 6% to 98% (M=61.27%, SD=22.46) and deep cohesion from 20% to 99% (M=79.93%, SD=18.1). Based on the data, most essays were coherent due to the use of linking devices that connect ideas throughout the essays. It represents students' knowledge of such phrases and their ability to apply them. The corpus was characterized by syntactic simplicity ranging from 2% to 72% (M=30.81%, SD=15.55). The reasons of this range may be differences between students' proficiency level and awareness of writing strategies.

The FKGL is a scale from 0 to 18 provided by the Coh-Metrix Easability Assessor. "The higher the number, the harder it is to read the text.

> READFKGL = (.39 x ASL) + (11.8 x ASW) - 15.59
> where:
> ASL = average sentence length = the number of words divided by the number of sentences. This is the same as READASL.
> ASW (comes from CELEX database) = average number of syllables per word = the number of syllables divided by the number of words." (Coh-Metrix version 3.0 indices)

The scale has three levels: basic, average, and skilled. Within all the three levels, there are two sublevels characterizing the readability level with a well-known book. The text complexity of the sample texts could be described with an average score of 8.56, which is the level of Harry Potter stories, according to the FKGL. Although Harry Potter stories may be lexically rich, their FKGL is in the average writing interval and does not rich the skilled text complexity level. Based on the one sample text, there was statistically highly significant difference between the mean score of sample essays and the minimum level of a skilled writing (from level 12 to 14), since the p-value was .000 ($p<.001$). However, there were five texts which reached the skilled level over score 14, one even achieved the level of an academic paper with 18. Results revealed a statistically significant negative correlation between students' syntactic simplicity scores and FKGL ($r=-.627$, $p=.000$).

Based on findings, students need assistance in improving their writing skills to focus on those crucial aspects that they need to develop, as Cotterall and Cohen (2003) also found. Such language skill development and familiarization with EAP vocabulary may be started in high school with those students who are planning to apply for English Studies. English specialization or extra

after-school activities would scaffold students' development in the target language to promote further education. Academic and technical vocabulary broadening should be focal points in their language improvement during their university studies.

Even though essays were created in the academic context, their vocabulary use resembled informal language and requires improvement. Nevertheless, the essays were written in the third week of their English studies; thus, students had time to improve their writing skills and overall language competences in the courses provided by the English Studies program during further semesters. The essay task played an essential role in estimating students' needs for language development. Research findings may contribute to the pedagogical development of academic essay writing at university in identifying deficiencies in students' mental lexicon and academic writing performance.

References

Bailey, S. (2011). *Academic Writing. A Handbook for International Students* (3^{rd} ed.) New York, NY: Routledge.

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes, 9*, 2-20. Retrieved on 24/03/2019.

Browne, C., Culligan, B. & Phillips, J. (2013). The New Academic Word List. Retrieved from `http://www.newgeneralservicelist.org`.

Browne, C., Culligan, B. & Phillips, J. (2013). The New General Service List. Retrieved from `http://www.newgeneralservicelist.org`.

Creme, P., & Lea, M. R. (2008). *Writing at university. A guide for students* (3rd ed.) New York, NY: Open University Press.

Cotterall, S., & Cohen, R. (2003). Scaffolding for second language writers: producing an academic essay. *ELT Journal, 57*(2), 158-166. Retrieved on 25/03/2019.

Csomay, E. (2005). Linguistic variation in the lexical episodes of university classroom talk. *Language in use. Cognitive and discourse perspectives on language and language learning. Georgetown University round table on languages and linguistics*, 150-162.

Duff, P. A. (2010). Language socialization into academic discourse communities. *Annual Review of Applied Linguistics, 30*, 169-192. Retrieved on 25/03/2019.

Dörnyei, Z. (2007). *Research Methods in Applied Linguistics.* Oxford: Oxford University Press.

Flowerdew, L. (2012). *Corpora and language education.* London: Palgrave Macmillan.

Luey, B. (2006). *Handbook for academic authors* (4^{th} ed.) New York, NY: Cambridge University Press.

Swales, J. (1993). *Genre analysis.* Great Britain, Glasgow: Cambridge University Press.

Academic reflective essay or anecdotal story writing: A study on pre-service EFL teaching portfolios

Viola Kremzer .
University of Pécs
Kremzer.viola@pte.hu

1 Introduction

Essays are widely used as assessment tasks in the academic context. Students are regularly required to submit academic essays as out-of-class and in-class assignments. A reflective essay is a type of academic text, in which writers critically reflect on their personal and professional development. In the case of teacher trainees, a reflective essay focuses on their experiences in connection with teaching and learning processes. Teacher trainees attend a number of academic seminars and lectures before submitting an EFL thesis and teaching portfolio. Thus, they become members of the academic community (Duff, 2010) and are able to write for academic audiences using English for Academic Purposes (EAP) and technical vocabulary related to the teaching profession. As Nation and Webb (2011) claim, topic-related technical vocabulary is crucial in academic writing to perform the academic functions in the text. University tutors scaffold student teachers' EAP vocabulary development so that they will be able to use the learnt expressions in academic discourse. On the other hand, the technical vocabulary of a particular field may range from a few hundred words to several thousand words, as Chung and Nation (2003) note. Therefore, depending on the subject area, the size of student technical vocabulary may vary.

Hungarian teacher trainees are required to submit a thesis and a portfolio according to the Higher Education Act of 15/2006. Pre-service and in-service teachers participating in the Teaching English as a Foreign Language (TEFL) program submit a portfolio at the end of their studies which is a collection of five documents written in English. The first document of the EFL teaching portfolio is a reflective essay discussing and summarizing the perceived professional development of the candidate. The second document is a lesson plan presenting a detailed design and schedule of a lesson that a teacher trainee developed and implemented during the teaching practice. In the final three documents candidates discuss three empirical studies that they conducted during their studies.

The present study aims at researching teacher trainees' essay writing skills and is embedded in research on English for academic purposes (Biber & Gray, 2010; Hamp-Lyons, 2011), more specifically academic essay writing (Bailey, 2011; Creme and Lea, 2008; Hinkel, 2001). According to Bailey (2011), an essay is written in an objective academic style. However, in the case of a reflective essay, it is inevitable to refer to the person of the writer, which creates a more personal and subjective tone of discourse. As Creme and Lea (2008) claim, the reflective essay is a "mixture of an academic and a professional approach" (p. 207) with personal experiences and reference to theory.

2 Corpus and methodology

To examine teacher trainees' academic writing skills in the Hungarian context, I conducted research guided by the following questions:
RQ1 How can the vocabulary of portfolio reflective essays written by teacher trainees be described in a corpus-based approach?
RQ2 What do syntactic measures indicate?
RQ3 What characterizes the cohesion and coherence of texts?
RQ4 What characterizes the narrativity of student essays?
RQ5 What readability level do the essays reach?

2.1 Corpus

The length of a reflective essay is not prescribed; it depends on the teacher trainee's decision and the advisor's recommendation on content and structure. Although the average text length was 759.8 words in the present study, reflective essays exceeded 2,000 words in several EFL teaching portfolios and consisted of three or four subsections. Sampling in the study was based on text length since the Coh-Metrix Easibility Assessor software can analyze texts below 1,000 words. The Essay Corpus complied from a sample of portfolios consisted of 20 reflective essays; 15,196 tokens altogether. The shortest text in the corpus was a 459-word, the longest one was a 986-word reflective essay. For ethical considerations, texts were coded in order not to reveal the person of the authors. Any information referring to the writers was also deleted from the texts (Dörnyei, 2007).

2.2 Methodology

In the present study, the corpus was analyzed in a quantitative approach. Tools of analysis were the VocabProfile VP-Compleat software program available on Tom Cobb's Compleat Lexical Tutor website and the Coh-Metrix Easibility Assessor. The essays were entered one by one into each program to create individual profiles as well as the complete Essay Corpus of the 20 texts to access the global vocabulary profile. Data were analyzed according for lexical sophistication, lexical density (LD), core general English vocabulary based on the New General Service List (NGSL; Browne, Culligan, & Phillips, 2013), academic vocabulary based on the New Academic Word List (NAWL; Browne, Culligan, & Phillips, 2013), and words not in any of these lists categorized as off-list.

Vocabulary analysis was followed by the examination of the 20 essays in the Coh-Metrix Test Easability Assessor, a text-analysis software available online calculating linguistic features; 1) narrativity, 2) syntactic simplicity, 3) word concreteness, 4) referential cohesion, 5) deep cohesion, and the 6) Flesch-Kincaid Grade Level (FKGL); of a text shorter than 1,000 words. These scores report on the readability profile of an essay that mirror possible fields to be improved. I used the IBM SPSS platform to analyze data obtained about vocabulary and readability profiles of the EFL teaching portfolio essays. The focus of SPSS analysis was to determine statistical differences, correlations and occurrent outliers.

3 Results

In the 20 sample reflective essays studied in the research, type/token ratios ranged between 0.37 and 0.5 (M=0.42, SD=0.04) and the lexical density figures between 0.48 and 0.58 (M=0.52, SD=0.02). Findings showed a high rate of general English vocabulary between 75.8%-87.9% (M=82.7, SD=3.33) and academic vocabulary ranging between 0.9%-2.9% (M=1.78%, SD=0.57). Browne, Culligan, and Phillips (2013) describe the coverage of an academic text written by native English speakers at school or university with approximately 6% from the NAWL. However, EFL learners encounter academic discourse thoroughly in higher education, at university. Since the authors of the sampled reflective essays were non-native English speakers, this percentage was not expected to cover their texts. There was a statistically highly significant difference (p=.000) between the 1.78% academic vocabulary coverage of the sampled texts and the 6% coverage found by Browne, Culligan, and Phillips (2013). Lexical analysis implies that teacher trainees need improvement regarding their writing skills, more specifically academic and technical vocabulary, which suggests the implementation of EAP vocabulary in course curriculums at universities.

Since texts focused on professional experiences, personal narratives featured the majority of the texts, which was also illustrated by the mean of 59.5% narrativity ratio ranging from 32% to 85% (SD=15.36). Even though reflective essays concentrate on the writer, begin academic texts, they ought not to be story-like. Thus, authors of reflective essays have to find the perfect balance between writing in academic discourse and implementing their own experiences. As the findings of the text analysis showed, writers could maintain the academic style by appropriate structures and framework. Thus, syntactic simplicity was characterized by scores ranging from 3% to 88% (mean=43%, SD=19.29). Range can be explained by differences in teacher trainees' writing skills. The ideas were well-connected in certain texts which was shown by a high level of coherence in those essays, still there was great range seen: referential cohesion between 9% and 77% (mean=48.05%, SD=20.69) and deep cohesion between 5% and 94% (mean=52.2%, SD=23.1). These numbers show that there were trainees who did not connect their ideas throughout the entire text to make it more understandable.

The FKGL is a scale provided by the Coh-Metrix Easibility Assessor program with levels ranging from 0 to 18. "The higher the number, the harder it is to read the text. [...]

> READFKGL = (.39 x ASL) + (11.8 x ASW) - 15.59
> where:
> ASL = average sentence length = the number of words divided by the number of sentences. [...]
> ASW (comes from CELEX database) = average number of syllables per word = the number of syllables divided by the number of words." (Coh-Metrix version 3.0 indices)

The scale has three levels: basic, average, and skilled; all three including two further sublevels defining with the readability with a well-known book. In the present study, the mean score

of the readability level of the sampled essays was 10.815, which was in the 6-to-12 interval describing the average readability level based on the FKGL. The texts may be linguistically rich; however, they did not reach the level of a skilled writing which would be above 12 according to the grading scale. There was no statistically significant difference (p>.01| p=.018) between the mean readability level of the sample and the minimum level (FKGL=12) of the skilled writing population. Apart from not reaching the level of a skilled writing, the sample texts did not reach the level of academic writing according to the FKGL which is in the interval between 14 to 16. Calculating with the minimum academic level of 14, statistically highly significant difference was obtained since the p-value was .000 in the one sample test. The results showed a statistically significant negative correlation between the texts' syntactic simplicity percentages and their FKGL (r=-.714, p=.000).

Findings showed that teacher trainees needed improvement concerning their academic mental lexicon since their texts resembled non-academic language use. The essays were well-constructed, but students seem to need scaffolding in academic writing to be able to address academic audiences more precisely. These findings could provide better insight into domains in need of improvement concerning writing reflective essays, such as academic vocabulary, syntactic structures, and readability levels. Thus, the present research may support writing skills courses in Hungary and other countries where English is taught as a FL in recognizing the need for more thorough implementation of academic vocabulary in syllabi. Results indicate a need for further research in the field to obtain a more elaborate picture of EFL teacher trainees' writing skills and competencies, as well as problem areas and needs for improvement.

References

Bailey, S. (2011). *Academic Writing. A Handbook for International Students* (3^{rd} ed.) New York, NY: Routledge.

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes, 9*, 2-20. Retrieved on 24/03/2019.

Browne, C., Culligan, B. & Phillips, J. (2013). The New Academic Word List. Retrieved from `http://www.newgeneralservicelist.org`.

Browne, C., Culligan, B. & Phillips, J. (2013). The New General Service List. Retrieved from `http://www.newgeneralservicelist.org`.

Chung, T. M., & Nation, P., (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language, 15*(2), 103–116.

Creme, P., & Lea, M. R. (2008). *Writing at university. A guide for students* (3rd ed.) New York, NY: Open University Press.

Cobb, T. (n.d.). The Compleat Lexical Tutor. Computer software. Available online at `www.lextutor.ca`.

Cotterall, S., & Cohen, R. (2003). Scaffolding for second language writers: producing an academic essay. *ELT Journal, 57*(2), 158-166. Retrieved on 25/03/2019.

Dörnyei, Z. (2007). *Research Methods in Applied Linguistics.* Oxford: Oxford University Press.

Duff, P. A. (2010). Language socialization into academic discourse communities. *Annual Review of Applied Linguistics, 30*, 169-192. Retrieved on 25/03/2019.

Hamp-Lyons, L. (2011). English for academic purposes. In E. Hinkel (Ed.) *Handbook of Research in Second Language Teaching and Learning, Volume 2* (pp. 89-105). New York: Taylor & Francis Group.

Hinkel, E. (2001). Giving personal examples and telling stories in academic essays. *Issues* In Applied Linguistics, 12(2), 149-170. Retrieved on 27/03/2019.

IBM Corp. (2015). IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY: IBM Corp.

Luey, B. (2006). *Handbook for academic authors* (4^{th} ed.) New York, NY: Cambridge University Press.

Nation, I. S. P., & Webb, S. (2011). Content-Based Instruction and Vocabulary Learning. In E. Hinkel (Ed.) *Handbook of Research in Second Language Teaching and Learning, Volume 2* (pp. 631-544). New York: Taylor & Francis Group.

Digital ou Numérique : un phénomène d'emprunt au cœur de la start-up nation ?

Lichao Zhu [1,2] et Gaël Lejeune [2]
[1] Textes Théories Numérique, Université Paris XIII
[2] Sens Texte Informatique Histoire, Sorbonne Université
lichao.zhu@gmail.com, gael.lejeune@sorbonne-universite.fr

Bien que l'anglais ne soit pas devenu la *Lingua Franca* de l'Internet, il est indéniable que son influence est très grande pour de nombreux domaines pour lesquels la circulation des connaissances se fait principalement par voie électronique [3]. Ceci se manifeste de nombreuses manières dont, pour ce qui intéresse la linguistique, les phénomènes d'emprunt et de calque. Nous nous intéresserons dans cet article à des phénomènes d'emprunts, d'adoption par une langue d'éléments langagiers provenant d'une autre langue.

Parmi les domaines pour lesquels l'influence de l'anglais est prégnante figurent en particulier la communication et l'informatique. Cette influence est visible sur la terminologie, on voit par exemple que dans le domaine de l'informatique le terme *implémenter*, emprunté à l'anglais *implement* supplante dans le français oral comme dans le français écrit le terme existant *implanter* pour désigner l'activité de mise en place d'un programme[1]. Il est intéressant de remarquer que certains usagers de la terminologie informatique jugent que l'on pourrait conserver les deux termes *implanter* et *implémenter* mais avec deux acceptions différentes[2].

La paire emprunt/terme supplanté qui nous intéresse pour cette étude est également issue de la terminologie informatique mais elle a quitté le domaine purement terminologique pour intégrer la langue courante. Il s'agit de la paire digital(e)/numérique, l'observation donc de l'utilisation de « digital », comme adjectif ou comme nom, en remplacement de « numérique » . La raison de notre intérêt pour cette paire est triple :

- On se situe à l'intersection de deux domaines (informatique et communication) dans lesquels les phénomènes d'emprunt, et particulièrement d'anglicismes, sont particulièrement foisonnants ;
- Le terme français « numérique » est répandu, facile à écrire et à prononcer et disposait de surcroît d'une certaine antériorité de sorte qu'il aurait pu être à l'abri de la supplantation par un emprunt ;
- Il s'agit d'un néologisme sémantique [5,6] puisque la forme « digitale » dans le sens « relatif aux doigts » est préexistante à son usage dans le sens de « numérique » ce qui n'est pas sans occasionner un certain nombre de réalisations langagières malheureuses (ou amusantes selon le point de vue où l'on se place).

Ce dernier aspect est particulièrement intéressant à étudier en diachronie, voire par exemple des travaux récents [4,2].

1. Voir par exemple sur le site de l'académie française un bref article sur le sujet : `http://www.academie-francaise.fr/implementer`
2. Voir par exemple : `http://jargonf.org/wiki/implémenter`

Digital a pour première acception « Qui a la forme d'un doigt » et « Relatif au doigt ; qui fait partie du doigt. » et trouve son étymologie en le mot latin impérial *digitalis* dont la signification est « qui a la grosseur d'un doigt » .

Son autre acception est « Qui est exprimé par un nombre, qui utilise un système d'informations, de mesures à caractère numérique. » et trouve son étymologie dans le langage informatique des années 1960 en anglais, en particulier dans l'unité lexicale « *digital computer* » . Le trésor de la langue française informatisé précise par ailleurs les relations entre ces deux significations : « digital » notamment dans digital computer « ordinateur digital » (du subst. digit « doigt » mais aussi « chiffre, [primitivement « compté sur les doigts »] »).

Tandis que *numérique* signifie « Qui concerne des nombres, qui se présente sous la forme de nombres ou de chiffres, ou qui concerne des opérations sur des nombres. » et « Qui désigne ou représente des nombres ou des grandeurs physiques au moyen de chiffres » .

Nous avons étudié l'usage des deux éléments de cette paire dans le corpus du journal le Monde de 1987 à 2017. La première grande vague d'utilisation de digital pour amener la notion de nombre se situe dans les années 1980-1990 autour des expressions « son digital » , « écran digital » et affichage digital. Ce qui n'est pas sans causer des incompréhensions pour les locuteurs puisque la plupart des occurrences de digital en tant qu'adjectif se retrouve dans « empreinte(s) digitale(s) » . Si le son est parfois « numérique » , l'affichage et l'écran ne le sont que très rarement. A partir de 2002 environ, l'usage de digital est presque systématique pour décrire ces réalités. En effet, pendant la même période, l'adjectif numérique se retrouve principalement associé à d'autres noms : photo, télévision et bouquet. Là encore, la différence dans le corpus est assez nette : « photo digitale » est très rare de même que « télévision digitale » et « bouquet digital » .

Au-delà des objets, telles que les décrivent les expressions citées ci-dessus, les processus arrivent rapidement au cœur des préoccupations exprimées par les journalistes. Or si la transformation est plutôt digitale, la fracture, elle, est surtout numérique. Ce phénomène est encore plus prégnant si l'on sort du corpus du Monde et de son écriture plus académique que d'autres supports : l'expression « fracture digitale » amène 8.000 résultats sur le moteur de recherche Google contre 2.000.000 pour « fracture numérique » . S'il faut bien sûr être prudent avec ce genre de tests, la différence d'ordre de grandeur semble très significative. D'ailleurs, si dans les articles sur la « transformation » , les auteurs prennent la peine de citer les deux adjectifs, c'est rarement vrai pour ceux concernant la « fracture » . On observe également que l'emprunt est nettement moins fréquent au pluriel, ce qui est peut être dû à des problèmes d'adaptation morphologique avec des noms masculins [1]. Au niveau de l'usage en tant que nom, il est intéressant de noter que le processus de « bas niveau » consistant à convertir des données dans un format traitable par un ordinateur reçoit le terme de « numérisation » beaucoup plus fréquemment que celui de « digitalisation » .

Nous présenterons un panorama plus large de ces usages en corpus, en comparant notamment les usages dans différents types de presse en ligne, dans le discours institutionnel (au sens politique) et les usages dans les forums et les documents de type « Présentation Power Point . Loin d'une vocation prescriptive, notre contribution viserait à présenter cette dualité dans une perspective tenant compte des publics visés et des types de discours (voire par exemple [7]).

Références bibliographiques

[1] Anna Anastassiadis-Syméonidis and Georgia Nikolaou. L'adaptation morphologique des emprunts néologiques : en quoi est-elle précieuse ? *Langages*, 183(3) :119–132, 2011.

[2] Emmanuel Cartier. Neoveille, système de repérage et de suivi des néologismes en sept langues. *Neologica : revue internationale de la néologie*, (10), July 2016.

[4] David Crystal. Language and the internet.*IEEE Transactions on Professional Communication* , 45(2) :142–144, June 2002.

[4] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)* , pages 1489–1501, Berlin, Germany, August 2016. Association for Computational Linguistics.

[5] Jacques Moeschler. Aspects de la néologie sémantique. *Langages* , 8(36) :6–19, 1974.

[6] Jean-François Sablayrolles. Extraction automatique et types de néologismes : une nécessaire clarification. *Cahiers de Lexicologie*, 1(100) :37–53, 2012.

[7] Xavier-Laurent Salvador. De quoi "numérique" est-il le nom dans la politique du monde moderne ? `https://www.lemondemoderne.media/numerique-et-politique-monde-moderne/`. Accessed : 2019-09-10.

Exploration des compétences langagières des enfants d'écoles maternelles en zone d'éducation prioritaire

Isabelle Rousset [1], Solange Rossato [2], Christine Lequette [3]et Elisabeth Latapie [4]
[1] LIDILEM, Université Grenoble Alpes
[2] LIG, Université Grenoble Alpes
[3] Médecin Conseil, Rectorat de l'académie de Grenoble
[4] DESDEN38, Rectorat de l'académie de Grenoble
isabelle.rousset@univ-grenoble-alpes.fr, solange.rossato@univ-grenoble-alpes.fr,christine.lequette@ac-grenoble.fr

1 Introduction

La direction départementale de l'éducation nationale de l'Isère (DSDEN38) a mis en place une formation des enseignants en Réseau d'Education Prioritaire (REP) autour d'ateliers langage à l'aide d'albums jeunesse de la TPS à la GS. Plusieurs études ([5,4,3,15,21]) ont montré le rôle de l'album dans le développement langagier des élèves de maternelle[1]. L'objectif des ateliers en groupe homogène est de proposer un cadre sécurisant permettant aux enfants de prendre la parole pour raconter un album qu'ils connaissent et ont travaillé en amont. La volonté de travailler avec des groupes homogènes s'inscrit dans le cadre de la pédagogie différenciée [18,20]. La question de l'évaluation du développement langagier s'est souvent posée dans un objectif de repérage précoce des troubles langagiers (IDE [9], le questionnaire Chevrie-Muller [6,2], et le BSEDS [1,26]). Ces outils sont utiles aux enseignants pour repérer les difficultés langagières des enfants (CDI [13]) mais souvent trop complexes. Dans ce cadre, nous avons élaboré une grille de positionnement adaptée pour permettre aux enseignants de constituer des groupes d'enfants homogènes. Pour valider cette grille de positionnement langagier, nous nous appuyons sur plusieurs travaux de recherche en acquisition du langage qui étudient le langage des enfants lors de suivis longitudinaux [11,16] ou de séances ponctuelles, en interaction avec des adultes. Ces dernières ont le plus souvent été proposées autour de supports vidéo (dessin animé) mais de récentes études s'appuient sur des albums jeunesse [23]. Les aspects temporels, tels que la longueur des prises de parole ou les débits articulatoires sont très souvent mesurés ([7,8,12,14,17,19,24,25,22]).
L'objectif de cette communication est de présenter la grille de positionnement, les premières passations de cette grille sur 750 élèves en REP. Nous présentons également l'analyse vidéo des ateliers langage d'une classe MS sur 3 mois en lien avec les niveaux obtenus sur la grille remplie par l'enseignante.

2 Méthodologie

Présentation de la grille de positionnement

Le point de départ de la réflexion était la grille validée de repérage enseignant du BSEDS 5-6 de la Santé Scolaire [26]. La grille de positionnement est organisée en 4 étapes de développement, chacune caractérisée par 6 items. Une étape est considérée comme validée lorsque 4 items sur

1. Ministère de l' Éducation nationale et de la Jeunesse, « Programme, ressources et évaluation - Mobiliser le langage dans toutes ses dimensions - Éduscol ». 2015 [En ligne]. Disponible sur : http ://eduscol.education.fr/cid91996/mobiliser-le-langage-dans-toutes-ses-dimensions.html#lien3.

6 sont acquis. La grille finale est présentée figure 1) Étant donné que cette grille est utilisée la TPS à la GS, une première étude consiste à vérifier que la progression du positionnement des enfants suit majoritairement l'âge et le niveau scolaire. Pour cela, 750 grilles remplies par des enseignant.e.s sur 8 écoles en REP (pour tous les élèves présents) ont été recueillies sur les mois de décembre et janvier de l'année scolaire 2015-2016.

Analyse vidéo lors d'ateliers langage

16 vidéos ont été recueillies par une étudiante de Sciences du Langage dans une classe de MS (REP+). Régulièrement présente, elle a filmé tous les ateliers langage de janvier à avril 2018 (3 albums différents). Pour les 18 enfants pour lesquels nous avons obtenus les autorisations, l'enseignante a rempli les grilles avant le début des enregistrements. Ces enfants ont des positionnements langagiers très disparates : 3 ont validé l'étape 1, 7 l'étape 2, 6 l'étape 3 et 2 enfants ont déjà validé l'étape 4, mais les groupes lors des ateliers langages ne sont pas homogènes. Les sessions concernent 3 à 5 enfants positionnés autour de l'enseignante et durent de 10 à 20 minutes. Les vidéos ont été annotées en tours de parole pour chaque locuteur grâce au logiciel ELAN [10]. Nous avons étudié les temps de parole cumulés de chaque enfant et de l'enseignant, leur nombre d'interventions et les durées de ces interventions pour chaque atelier langage.

3 Premiers Résultats

Grille de positionnement, âge et niveau scolaire

A partir des 750 grilles remplies par des enseignant.e.s, nous avons calculé la moyenne des étapes validées pour tous les enfants de même âge (en mois). Les résultats sont présentés figure 2 et montrent une forte corrélation (R^2=0.76). Si la tendance globale est clairement marquée, l'étude de l'étape validée pour chaque enfant montre une grande disparité existante au sein d'un même niveau scolaire (figure 3). En effet, on observe en PS de grands écarts de niveau langagier avec 10% des enfants qui ne valident pas l'étape 1, 15% à l'étape 1, 20% à l'étape 2, 44% à l'étape 3 et 11% à l'étape 4. La proportion d'élèves ayant validé l'étape 4 progresse en MS (47%) pour atteindre 70% des élèves de GS tandis que 25% d'entre eux sont encore à l'étape 3. Devant de tels écarts de langage, le travail en petits groupes de niveau homogène apparaît clairement nécessaire pour faire progresser tous les élèves.

Prises de parole lors d'ateliers langage

Pour étudier les prises de parole lors de l'atelier langage, nous avons calculé pour chaque individu le pourcentage du temps de parole, le pourcentage du nombre d'interventions et la durée moyenne des interventions par atelier. Les résultats sont moyennés en fonction du niveau langagier et présentés Table 1. Ces mesures montrent des progressions claires entre l'étape 1, les étapes 2 et 3, et l'étape 4. Peu de différences émergent entre les étapes 2 et 3 malgré une légère augmentation de la durée moyenne des interventions. Cette analyse confirme l'hétérogénéité des profils langagiers au sein d'un même niveau scolaire.

4 Discussion

Nous avons montré que la grille de positionnement permet de situer l'enfant au sein d'une progression langagière qui est corrélée avec l'âge (en mois) des enfants de la TPS à GS. Un autre apport de l'analyse des grilles remplies par les enseignant.e.s a permis de mettre en évidence

Outil repérage : 4ème étape (MS/GS)	Oui	Non
Intervient pour des prises de parole dans le groupe classe		
Prononciation correcte		
Utilisation systématique des déterminants et des pronoms		
Utilisation des phrases complexes avec propositions subordonnées		
Emploi de temps variés		
Conversation proche de celle d'avec un adulte		

Outil repérage : 3ème étape (PS/MS/GS)	Oui	Non
Intervient verbalement autrement que par oui ou non lorsqu'il est interrogé		
Intervient pour des prises de parole dans le groupe classe		
Utilise le « je »		
Utilisation de déterminants et de pronoms (on accepte les erreurs en genre et nombre)		
Utilisation des phrases complexes (GN GV + complément(s) de phrase)		
Emploi plusieurs temps : présent - passé		

Outil repérage : 2ème étape (TPS/PS/MS)	Oui	Non
Est compréhensible en français ou dans sa langue maternelle		
Intervient verbalement autrement que par oui-non lorsqu'il est interrogé		
Réutilise les formules données en classe (ex : j'ai fini, chacun son tour, ...)		
Fait des phrases avec au moins 3 mots dans l'ordre		
Suit des consignes simples données à l'oral		
Emploi plusieurs temps : présent - passé (on accepte encore les erreurs "il a prendu")		

Outil repérage : 1ère étape (TPS/PS)	Oui	Non
Sait se faire comprendre, en français dans sa langue maternelle et/ou par gestes		
Suit des consignes simples données à l'oral		
Communique par des mots isolés		
Répète spontanément des mots ou groupes de mots		
Combine des mots "gâteaux - encore"		
Emploi des verbes		

FIG. 1 : Grille de positionnement langagier remplie par les enseignant.e.s

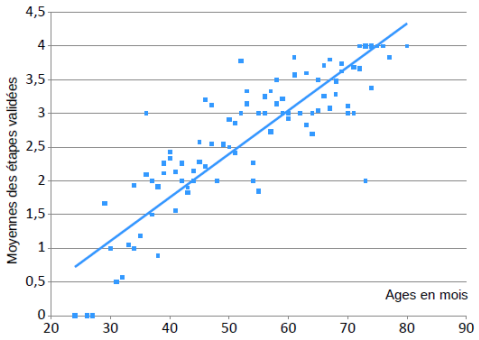


FIG. 2 : Moyenne des étapes validées en fonction de l'âge en mois.

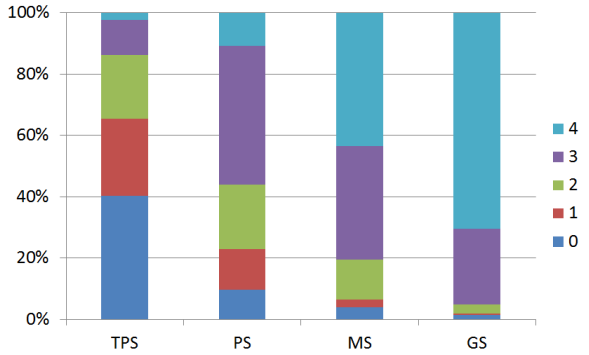


FIG. 3 : Répartition des enfants dans chaque étape de la grille en fonction de leur niveau scolaire : TPS (62 élèves), PS(225), MS(191), GS(196)

Niveau langagier	E1	E2	E3	E4	Enseignante
Nombre de personnes concernées	3	7	6	2	1
Moyenne des pourcentages de temps de parole (en %)	3	9	7	20	68
Moyenne des pourcentages d'interventions (en %)	6	14	13	26	51
Moyennes des durées d'interventions (en s)	0,98	1,37	1,54	2,02	3,47

TAB. 1 : Analyse des prises de parole en fonction du niveau langagier

l'hétérogénéité des niveaux langagiers dans un même niveau scolaire. L'analyse plus détaillée des vidéos des ateliers langages dans une classe de MS a également illustré cet écart, avec des enfants qui parlent 3% du temps tandis que d'autres atteignent 20% du temps de parole lors d'un atelier langage. Nous observons que les différences des mesures temporelles entre les étapes 2 et 3 sont minimes. La prochaine étape est de compléter ce travail par des analyses linguistiques explorant le contenu des interventions langagières des enfants. Nous souhaitons également comparer ces résultats obtenus sur des groupes hétérogènes avec l'analyse d'ateliers langage avec des groupes homogènes, permettant une meilleure répartition de la parole. Enfin, on remarque que même dans un atelier langage dont l'objectif est de faire raconter aux enfants un album, la parole de l'enseignant reste prépondérante, rejoignant ainsi les observations de [23]. Une autre perspective de travail étant de comparer différences approches permettant de réduire l'intervention verbale de l'enseignant.e.

5 Références bibliographiques

[1] V. Azzano, M. Jacquier-Roux, D. Lepaul, C. Lequette, G. Pouget, and M. Zorman. Bilan de Santé Évaluation du Développement pour la Scolarité à 5/6 ans, 2011.

[2] P. Boisseau. *Enseigner la langue orale en maternelle.* Retz-CRDP de Versailles, Paris, 2005.

[3] E. Canut. L'apprentissage du langage oral à l'école maternelle : rôle, modalités et enjeux des interactions langagières entre adulte et enfant, 2007. 00002.

[4] E. Canut, F. Bruneseaux-Gauthier, and M. Vertalier. *Des albums pour apprendre à parler : les choisir, les utiliser en maternelle.* CRDP de Lorraine, 2012.

[5] E. Canut, C. Masson, and M. Leroy. *Accompagner l'enfant dans son apprentissage du langage : De la recherche en acquisition à l intervention des professionnels.* Hachette Éducation, Apr. 2018.

[6] C. Chevrie-Muller and J. Goujard. Validation d'une méthode de dépistage précoce des troubles du langage. *Approche neuropsychologique des apprentissages chez l'enfant*, 2(1) :30–9, 1990.

[7] J.-M. Colletta, C. Pellenq, A. Hadian-Cefidekhanie, and I. Rousset. Developmental changes in articulation rate and phonic groups during narration in French children aged four to eleven years. *Journal of Child Language*, 45(06) :1337 –1356, Nov. 2018.

[8] J.-M. Colletta, C. Pellenq, and I. Rousset. Évolution du débit de parole chez l'enfant francophone dans des tâches narrative et conversationnelle. In *Actes des XXVIèmes Journées d'Étude sur la Parole,* Avignon, France, 2008.

[9] M. Duyme and C. Capron. L'inventaire du développement de l'enfant (ide). normes et validation françaises du child development inventory (cdi). *Devenir*, 22(1) :13–26, 2010.

[10] M. P. I. for Psycholinguistics. ELAN (Version 5.4), 2019.

[11] K. D. Hall, O. Amir, and E. Yairi. A longitudinal investigation of speaking rate in preschool children who stutter. *Journal of speech, language, and hearing research : JSLHR,* 42(6) :1367–1377, Dec. 1999.

[12] C. Hulme, N. Thomson, C. Muir, and A. Lawrence. Speech rate and the development of short-term memory span. *Journal of Experimental Child Psychology*, 38(2) :241–253, Oct. 1984.

[13] S. Kern, J. Langue, P. Zesiger, and F. Bovet. Adaptations françaises des versions courtes des inventaires du développement communicatif de MacArthur- Bates. *ANAE-Approche Neuropsychologique des Apprentissages chez l'Enfant*, (107/108) :217–228, 2010.

[14] F. J. Koopmans-van Beinum. Cyclic effects of infant speech perception, early sound production, and maternal speech. In *Proceedings of the Institute of Phonetic Sciences*, volume 17, pages 65–78, 1993.

[15] R. Léon. *La littérature de jeunesse à l'école : pourquoi ? comment ?* Hachette Éducation, Paris, nouv. edition edition, 2004.

[16] A. Morgenstern. *L'enfant dans la langue.* Presses Sorbonne Nouvelle, Paris, Nov. 2009.

[17] I. S. B. Nip and J. R. Green. Increases in Cognitive and Linguistic Processing Primarily Account for Increases in Speaking Rate With Age. *Child Development*, 84(4) :1324–1337, July 2013.

[18] P. Perrenoud. *La pédagogie à l'école des différences : Fragments d'une sociologie de l'échec.* ESF Editeur, Issy-les-Moulineaux France, 4e éd edition, Oct. 2005.

[19] R. H. Pindzola, M. M. Jenkins, and K. J. Lokken. Speaking Rates of Young Children. *Language Speech and Hearing Services in Schools,* 20(2) :133, Apr. 1989.

[20] C. Ponce. Pédagogie différenciée. *Revue française de pédagogie*, 114(1) :97–102, 1996.

[21] N. Prince. *La littérature de jeunesse en question(s)*. PU Rennes, Rennes, 2009.

[22] P. Péroz. Allongement des prises de parole et apprentissage du langage à l'école maternelle. May 2005.

[23] P. Péroz. Langage oral : la pédagogie de l'écoute, Oct. 2017.

[24] B. P. Ryan. Speaking rate, conversational speech acts, interruption, and linguistic complexity of 20 pre-school stuttering and non-stuttering children and their mothers. *Clinical Linguistics & Phonetics*, 14(1) :25–51, 2000.

[25] J. F. Walker and L. M. D. Archibald. Articulation rate in preschool children : a 3-year longitudinal study. *International Journal of Language & Communication Disorders*, 41(5) :541–565, 2006.

[26] M. Zorman and M. Jacquier-Roux. BSEDS 5-6 Un dépistage des difficultés de langage oral et des risques de dyslexie qui ne fait pas l' économie de la réflexion clinique. *A.N.A.E. (Approche Neuropsychologique des Apprentissages chez l'Enfant)*, Dépistage des troubles de l'apprentissage scolaire :tests, bilans, batteries ; intérêt et limites(66) :48–55, 2002.

Démonstrations

Graph Matching for Corpora Exploration

Bruno Guillaume
Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
Bruno.Guillaume@inria.fr

1 Introduction

When annotated corpora are available, it is possible to conduct more fined-grained corpus linguistics studies. In this paper, we will focus on two kinds of corpora: syntactically and semantically annotated corpora.

When dealing with syntax, most linguistic models are based on tree structures. The two mainstream syntactic traditions, Phrase Structure and Dependencies, propose to represent the syntax of a sentence as a tree. Nevertheless, if we have a closer look to popular corpora, they introduced mechanisms that are going beyond tree structure. The figure 1 on the left shows a subpart of the first example given in Stephen Clark's presentation *Penn Treebank Parsing*[1] (Taylor *et al.*, 2003). There are additional (dotted) edges which describe the links between some traces and their antecedent. If we consider that these links are in the structure, it is no longer a tree but a graph. Similarly, the figure 1 on the right shows the first example given in the documentation page of the Prague Dependency Treebank[2] (Hajič *et al.*, 2017). The blue arrow links two nodes with a common ancestor and so the structure is a graph.

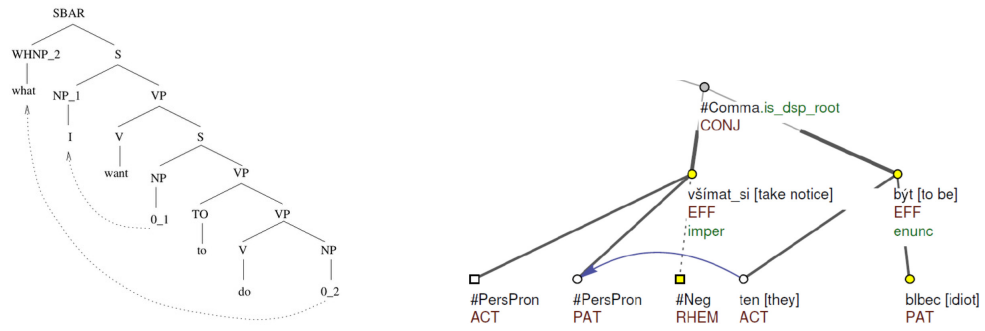


Figure 1: Examples of structures from the Penn Treebank and the Prague Dependency Treebank

When semantics is taken into account, most formalisms rely either an explicit notion of graphs or on other mathematical structures (logical formulae or λ-terms) that can be represented as graphs.

These examples drive us to consider all these structures as graphs and to propose GREW (Bonfante *et al.*, 2018), a graph-based tool to deal with them.

1. http://www.cs.ox.ac.uk/files/552/stat_parse2.pdf
2. https://ufal.mff.cuni.cz/pdt3.0/documentation

2 Graph Matching

We consider an usual mathematical definition of a *graph*: a set of *nodes* and a set of relations between the nodes (called *edges*). To cope with Natural Language Processing specificities, we add feature structures on nodes (to handle phonological forms, lemmas, POS, morphological features…) and labels on edges to be able to type relations.

Graph matching is also a mathematical notion which is well defined: it is the process of searching for a *pattern* (which is expressed itself as a graph) and finding all possible ways to recognize the pattern in the host graph. We call each solution of the search of the pattern an *occurrence*.

In GREW, patterns are written in a dedicated syntax with three kind of constraints:

- Global constraints about the whole graph (testing for the presence of cycles or for a tree structure),
- Positive constraints where the user describes a set of nodes, of relations and constraints on them,
- Negative constrainsts which are used as filters on the positive constraint output.

All these parts are optional and negative constraints can be repeated. If there are more than one negative constraints, they are intrepreted as independant filters. Some examples of each constraint are given below.

For a convenient usage, an online web application (named GREW-MATCH[3]) is available. The user selects a corpus (a few hundreds are proposed) and writes a graph pattern. The application returns the numbers of occurrences found in the corpus and displays some of the results. A tutorial mode is available to help new users to learn the concrete syntax of patterns.

3 Application to syntactically annotated corpora

GREW-MATCH is available on all Universal Dependencies (UD) (Nivre *et al.*, 2016) corpora and the examples below are done on one of them: the UD_ENGLISH-GUM (Zeldes, 2017) which contains 4,399 annoted trees and 80 176 tokens (in UD version 2.3).

Our first example computes some statistics about the `conj` relation which is used in UD to link heads of conjucts in a coordination construction. On the UD_ENGLISH-GUM corpus, we observe 2,563 occurrences of `conj` relations with two homogenoeous conjuncts (first pattern below) and 535 occurrences of coordination of unlikes (second pattern). Among the 535, the most productive pair of POS is when `C1` is adjective whereas `C2` is a verb (third pattern) with 71 occurences. An example of the last case is *They are simply afraid and hate modernity.* where the `conj` relation links *afraid* and *hate*.

```
pattern { C1 -[conj]-> C2 ; C1.upos = C2.upos }
pattern { C1 -[conj]-> C2 ; C1.upos <> C2.upos }
pattern { C1 -[conj]-> C2 ; C1[upos=ADJ]; C2[upos=VERB] }
```

3. `http://match.grew.fr`

With some other patterns, we can examine the most frequent POS of conjucts in the 2,563 occurrences: there are nouns (1,102), verbs (868), proper names (396), adjectifves (187).

Our second example explores different kinds of noun phrases (NP). We would like to estimate the proportion of NP built with or without a determiner and to see if these proportion varies depending of the syntactic role of the NP in the sentence. We obseve that a majority (59.3%) of the nouns of the corpus are used without any determiner. The table 1 gives these proportion for some syntactic role of NP.

	subj	**obj**	**obl**	**nmod**	**compound**	**other**	**Total**
with	1,070	1,412	1,470	1,047	8	1,043	6,050
det	50.9%	48.7%	50.4%	43.3%	0.5%	35.2%	40.7%
without	1,031	1,488	1,447	1,371	1,554	1,922	8,813
det	49.1%	51.3%	49.6%	56.7%	99.5%	64.8%	59.3%
Total	2,101	2,990	2,917	2,418	1,562	2,965	14,863

Table 1: Noun phrases with or without determiner

Patterns with negative constrains are used to get some numbers of the table. For instance, the 1,488 occurrences of direct object nouns without determiner are found with the pattern:

```
pattern { N[upos=NOUN]; * -[obj]-> N;}
without { N -[det]-> *; }
```

4 Application to semantically annotated corpora

We have chosen the Abstract Meaning Representation (Banarescu *et al.*, 2013) (AMR) because it provides freely available annotated corpora but the method may be applied to any semantic framework as soon as the annotated structures are converted into graphs that can be read by the tool.

From the AMR website[4], two corpora are available: the translation in English of the Saint-Exupéry's novel *The Little Prince* (1,562 sentences) and the Bio AMR Corpus (6,452 sentences from 3 full cancer-related PubMed articles). Both corpora are available on GREW-MATCH but we use here *The Little Prince* where sentences are much shorter and simpler. In this section, in order to spot the potential usages of the tool, we focus only on examples where the graph structure is essential.

In AMR, relations between predicates and arguments are encoded with labels `ARG0` (roughly for the semantic agent) and `ARG1`, `ARG2`, …, `ARG5` for other predicative roles. The following pattern describes constructions where two predicates `P1` and `P2` share the same `ARG0` argument (identified by `N`) and such that `P2` is an argument of `P1`. Note that it is not possible to search for this kind of structure if the pattern itself cannot have a graph structure (because if the sharing of `N`).

```
pattern { P1 -[ARG0]-> N; P2 -[ARG0]-> N; P1 -[ARG1|ARG2|ARG3]-> P2 }
```

4. `https://amr.isi.edu`

141 occurrences of the pattern are found in the corpus. Two of them are shown in the figure 2 (these figures are also examples of the graphical output avalaible in GREW-MATCH where parts of the graph corresponding to the pattern image are highlighted).

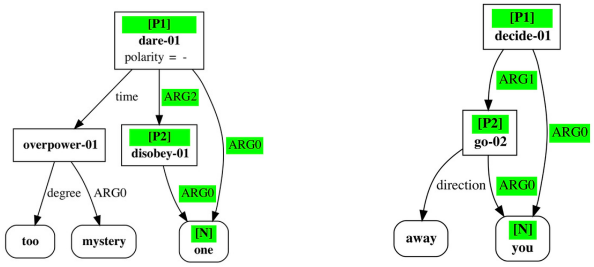


Figure 2: Examples of AMR structures found with GREW-MATCH

One can also obtain some statistics about the occurrences, for instance about the predicates involved in this construction. In the previous example, around 50 different predicates are used in the `P1` node for this construction. The most frequent predicate is by far `say-01` (23 occurrences); the next ones being `begin-01` (7 occurrences), `continue-01`, `try-01`, `want-01` (each with 6 occurrences).

As say earlier, GREW-MATCH allows also to search for global properties on the graphs. In AMR guidelines[5], it said that: "*Approximately 0.3% of AMRs are legitimately cyclic*"; but we can observe that it is underestimated at least for the two available corpora. With the pattern `global { is_cyclic }`, we found 35 sentences with AMR containing a cycle (2.24% of the 1,563 sentences). With patterns on nodes; we can explore further these cycles (6 are of length 2; 26 of length 3 and 8 of length 4). In the Bio AMR Corpus, 2.71% of the graphs are cyclic.

5 Conclusion

The tool we propose can be used in many different ways to explore the graph structure of annotated linguistic corpora. Is is currently used on syntax and on semantics but usages on other kinds of corpora can be imagined. In fact, the method can be used on any graph structure and for instance, it has been recently used on lexical databases.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, Nathan Schneider (2013). Abstract meaning representation for sembanking. *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

Guillaume Bonfante, Bruno Guillaume, Guy Perrier (2018). Application of Graph Rewriting to Natural Language Processing. *ISTE Wiley*, 1, pp.272, Logic, Linguistics and Computer Science Set

Jan Hajič, Eva Hajičová, Marie Mikulová, Jiří Mírovsky (2017). Prague dependency treebank. *Handbook of Linguistic Annotation*, pages 555–594. Springer.

5. `https://github.com/amrisi/amr-guidelines/blob/master/amr.md#cycles`

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, *et al.* (2016). Universal dependencies v1: A multilingual treebank collection. *Proceedings of LREC 2016*, pages 1659–1666.

Ann Taylor, Mitchell Marcus, Beatrice Santorini (2003). The Penn Treebank: An Overview. pages 5–22, Springer Netherlands, Dordrecht.

Amir Zeldes (2017). The GUM corpus: Creating multilayer resources in the classroom. *Resources and Evaluation*, 51(3):581–612.

Nouvelles fonctions logicielles pour l'analyse de grands corpus

Philippe Martin
Laboratoire LLF, UFRL, Université de Paris
philippe.martin@utoronto.ca

1 Introduction

Le logiciel WinPitch d'analyse acoustique de la parole (WinPitch, 2019), à la différence de systèmes plus connus utilisant des scripts, intègre dans sa dernière version différentes fonctions intégrées facilitant et étendant les possibilités d'analyse de la parole, que ce soit parole de laboratoire ou de grands corpus oraux spontanés. Ces nouvelles fonctions permettent 1) l'annotation prosodique d'enregistrements très bruités et multi-sources (ex. analyse de cris de primates), 2) la segmentateur automatique par alignement de parole de synthèse en 36 langues, 3) l'annotation automatique des syllabes accentuées en français, 4) l'analyse des correspondances texte-parole par concordancier intégré pour l'analyse de grands corpus oraux, 5) l'affichage simultané des courbes de pression sous glottique, nasale et intra-orale et 6) l'analyse de signaux EEG.

2 Annotation prosodique d'enregistrements très bruités

Certains enregistrements de parole, en particulier de parole spontanée, défient toute analyse fiable de la fréquence fondamentale quel que soit l'algorithme utilisé (Martin, 2019). En superposant un spectre de Fourier à bande étroite, de manière à visualiser les harmoniques et en particulier la fondamentale (ou la deuxième harmonique si la première n'est pas visible sur le spectre), un annotateur peut établir graphiquement à la souris avec des commandes ergonomiques une annotation par contour ou par niveaux épousant les mouvements mélodiques visibles sur le spectre de Fourier (figure .1).

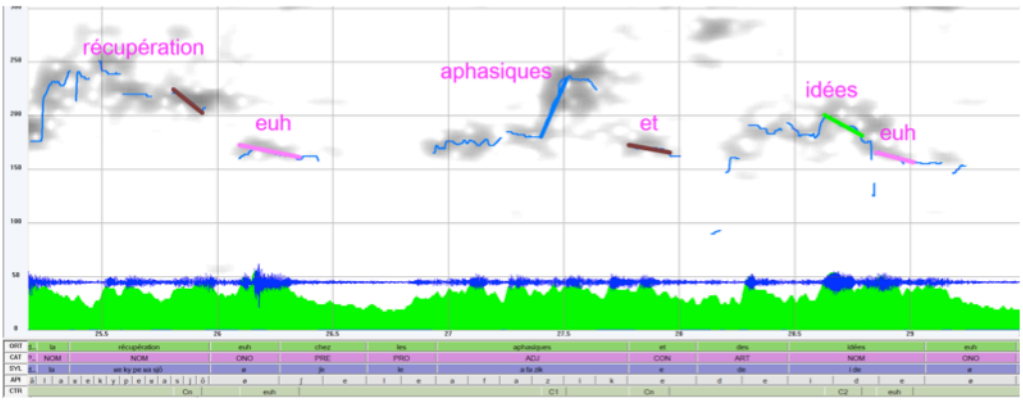


FIG. 1 : Annotation prosodique graphique d'enregistrements très bruités par segments linéaires

La catégorisation de ces mouvements peut de plus être établie automatiquement à partir de certains critères définis préalablement par l'utilisateur (par exemple durée vocalique, valeur du glissando, hauteur mélodique, etc.).

3 Segmentateur par alignement de parole synthétique (36 langues disponibles)

À la différence des systèmes existants basés sur des similitudes spectrales avec des modèles de mélange de gaussiennes qui exigent un apprentissage et la phonétisation du texte (ex. : EasyAlign, 2019), la segmentation en mots et en phones effectuées à partir d'une transcription orthographique est réalisée par alignement forcé de phrases synthétisées pour chaque segment de transcription entre deux pauses (Malfrère & Dutoit, 1997). Les avantages de cette méthode sont, entre autres a) l'intégration des liaisons, enchainements et autres phénomènes contextuels par le synthétiseur (fonction particulièrement importante en français), (figure .2) et b) l'application multilingue due à l'intégration dans WinPitch de synthétiseurs texte-parole disponibles sous Windows 10 en français (France, Québec, Suisse), néerlandais (Belgique, Pays-Bas), anglais (britannique, États-Unis, Canada, Australie, Inde), allemand (Allemagne, Autriche, Suisse), russe, finlandais, suédois, norvégien, danois, italien, espagnol (Espagne, Mexique), catalan, portugais (Portugal, Brésil), polonais, slovène, slovaque, tchèque, roumain, bulgare, grec, hongrois, indonésien, croate, malais, hébreu, tamil, hindi, turc, thaï, vietnamien, coréen, japonais, mandarin et cantonais.

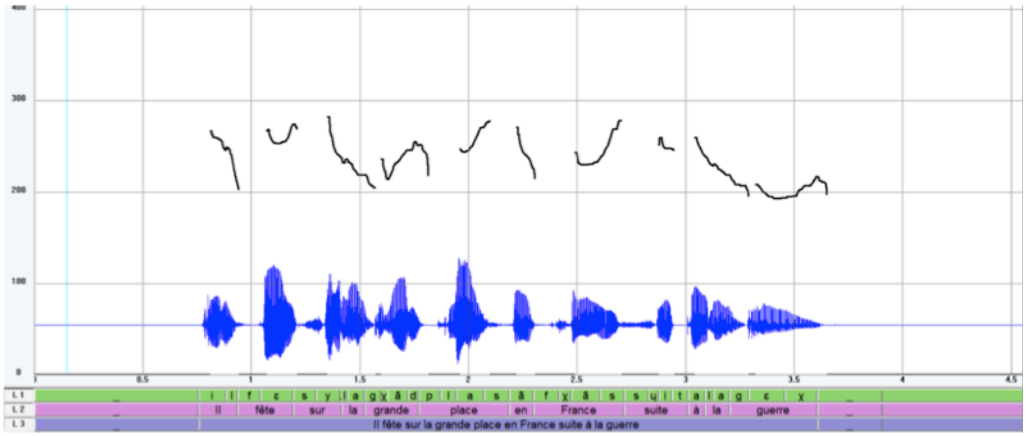


FIG. 2 : Segmentation automatique par alignement de parole synthétique du texte (exemple en français québécois)

4 Annotation des syllabes accentuées en français

Le français étant une langue dépourvue d'accent lexical, l'annotation des syllabes accentuées, hors emphase, se révèle problématique pour les annotateurs (Christodoulides & Avanzi, 2014).

Un algorithme nouveau, opérant à la fois en mode montant (bottom-up) et descendant (top-down) a été implémenté en intégrant les mécanismes de perception de l'accent, basés à la fois sur des critères acoustiques et morphologiques. Parmi les critères retenus, a) présence d'une pause de plus de 250 ms après une syllabe accentuée, b) variation mélodique supérieure au seuil de glissando (Rossi, 1971), c) accentuabilité des catégories lexicales (verbe, adjectif, adverbe et nom), d) critères rythmiques (distance minimale de 250 ms et maximale de 1250 ms entre deux syllabes accentuées successives). Les comparaisons avec les annotations manuelles disponibles

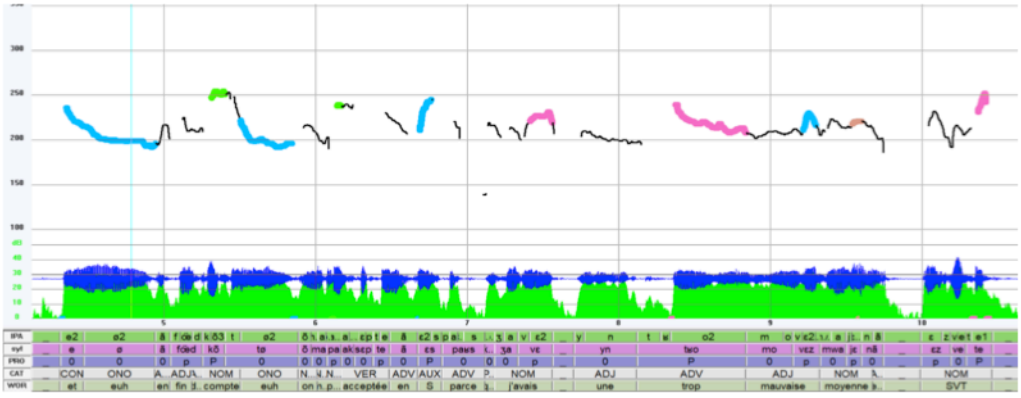


FIG. 3 : Annotation des syllabes accentuées en français (comparaison annotation manuelle et automatique –les contours mélodiques sont surlignés en différentes couleurs selon la classe de l'accent syllabique). Fichier D103, corpus Rhapsodie).

dans Rhapsodie (2019) montrent un niveau important de correspondance, mais montrent aussi les limites de l'annotation manuelle, influencée entre autres par le débit de parole habituel des annotateurs (figure .3).

5 Concordancier intégré pour l'analyse de grands corpus

WinPitch intègre in concordancier permettent de retrouver quasi instantanément un segment de parole à partir d'une entrée orthographique, avec l'affichage des données acoustiques pertinentes (courbes d'intensité, de fréquence fondamentale, spectrogramme, etc.). L'affichage de ces données (plus de 4000 concordances peuvent être affichées en quelques secondes, figure .4) sur des données disponibles en format *textgrid* (Praat, 2019) ou *trs* (Transcriber, 2019), ce qui permet d'établir très rapidement des corrélations entre item syntaxique ou lexical (par exemple *enfin, selon,* etc.) et les variations mélodiques correspondantes (qui auparavant demandaient parfois des mois de travail d'édition).

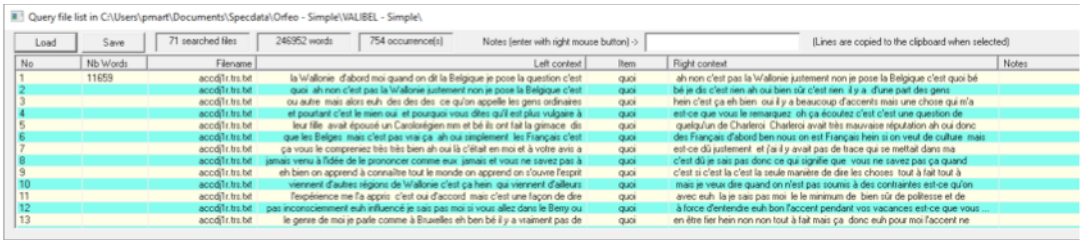


FIG. 4 : Concordancier intégré pour l'analyse de grands corpus. Dans cet exemple le segment de parole correspondant, ainsi que les courbes d'intensité, de fréquence fondamentale et le spectrogramme sont affichés en cliquant sur la ligne sélectionnée pour le mot *quoi* avec 754 entrées (Corpus Valibel)

6 Affichage des courbes de pression sous-glottique, intra-orale et des débits d'air buccal et nasal

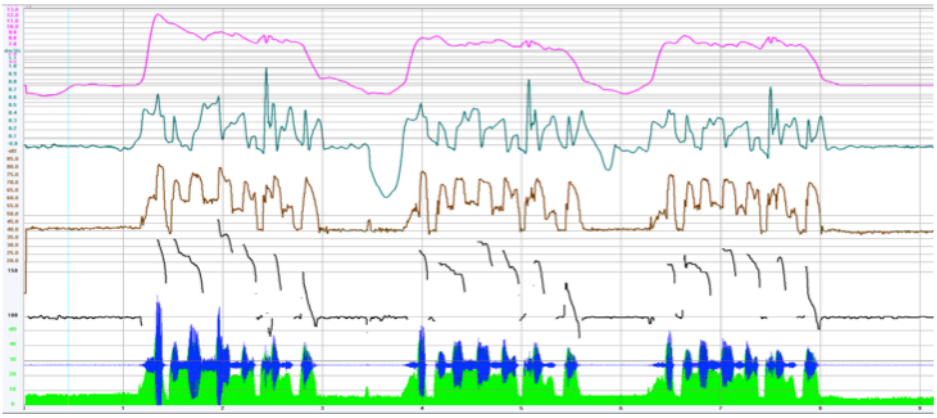


FIG. 5 : Affichage simultané des courbes de pression sous-glottique, de débit d'air buccal et nasal, ainsi que l'intensité et la fréquence fondamentale

N'importe quelle courbe au format RIFF peut être affichée, dans la couleur désirée, en même temps que les courbes oscillographique, d'intensité et de fréquence fondamentale d'un signal de parole. Il est aussi possible d'aligner temporellement les différentes courbes qui auraient été obtenus par des instruments aux caractéristiques distinctes. Les données relatives à chaque courbe peuvent être numérisées et obtenues sur un tableur (Excel par exemple) en un simple clic (figure .5).

7 Analyse des signaux EEG

L'analyse des potentiels évoqués ouvre de nouvelles possibilités en recherche sur la parole, que ce soit en perception ou en production. WinPitch intègre la plupart des fonctions qui requièrent habituellement l'utilisation du logiciel Matlab, ainsi que des systèmes associés (par exemple EEGLAB), mais avec un temps de calcul très inférieur et une ergonomie plus conviviale. Parmi les fonctions disponibles a) mapping des électrodes (jusqu'à 256 canaux), b) affichage des moyennes par essai et par région d'électrodes, c) affichage simultané du spectre de Fourier et par ondelettes avec la fréquence fondamentale et l'intensité du signal de parole, d) analyse par composantes indépendantes (ICA), et cohérence de phase inter-essai (ITC), etc. (figure .6).

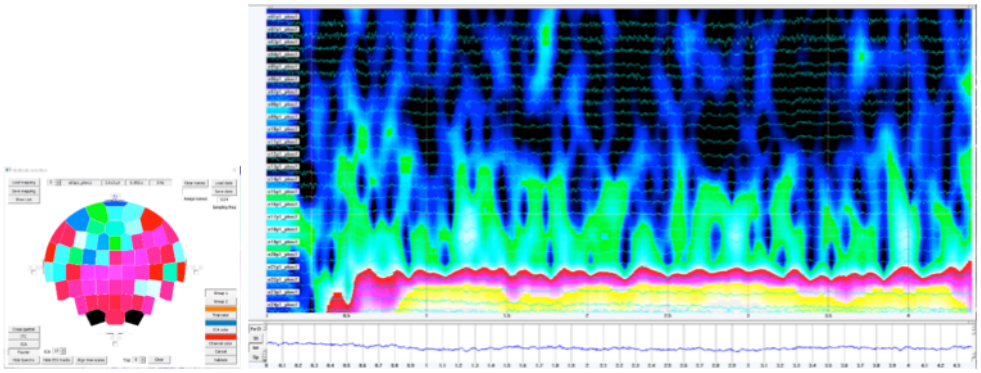


FIG. 6 : Analyse de signaux EEG. Analyse spectrale à partir des signaux d'un casque EEG à 61 électrodes et mapping des différents canaux

Références bibliographiques

Christodoulides, G. & Avanzi, M. (2014) An Evaluation of Machine Learning Methods for Prominence Detection in French. *Proc. Interspeech 2014*, 116-119.

EasyAlign (2019)`http://latlcui.unige.ch/phonetique/easyalign.php`

Malfrère F. & Dutoit Th. (1997) High Quality Speech Synthesis for Phonetic Speech Segmentation, *Proceedings EuroSpeech'97,* 2631-2634.

Martin, Ph. (2019) Tools for fundamental frequency estimation in Rhapsodie : in *Rhapsodie : A prosodic and syntactic treebank for spoken French* (SCL 89), Lacheret, A, Kahane, S. and Pietrandrea, P. ed., Amsterdam : Benjamins, 261-269.

Praat (2019) `https://www.praat.org`

Rossi, M. (1971) Le seuil de glissando ou seuil de perception des variations tonales pour la parole. *Phonetica*. n° 23, 1971, 1-33.

Transcriber (2019) `https://transcriber.fr.softonic.com/`

WinPitch (2019) `http://www.winpitch.com`

Learning Business English: A preliminary analysis of an Italian ESP Learner Corpus

Anna Romagnuolo , Claudio Latini et Mirko Meloni
CML Research Unit[1], University of Tuscia, Viterbo, Italy
romagnuolo@unitus.it, mirkomeloni1@gmail.com, claudio.lat95@gmail.com

1 Introduction

The paper will present the functioning of a *Visual Basic for Applications* Software (VBA) purposely designed by Tuscia University CML research group to analyze an ESP Learner Corpus made of 400 Business writing exam tests written by Italian University students during the final test of a Business English Writing Course and gathered in the years 2010-2013. The illustration of the software functions will also show the preliminary results of the analysis of the corpus collection, manually annotated according to the Louvain error tagging taxonomy, which has been purposely modified for the investigation of pragmatic errors, besides grammar mistakes. The ultimate objective of the software use is the synchronic and diachronic observation of Italian Learners' most frequent errors often caused by mother-tongue interference, cross-cultural communication differences and students' lack of professional skills and hands-on experience of the business world.

2 Corpus and software description

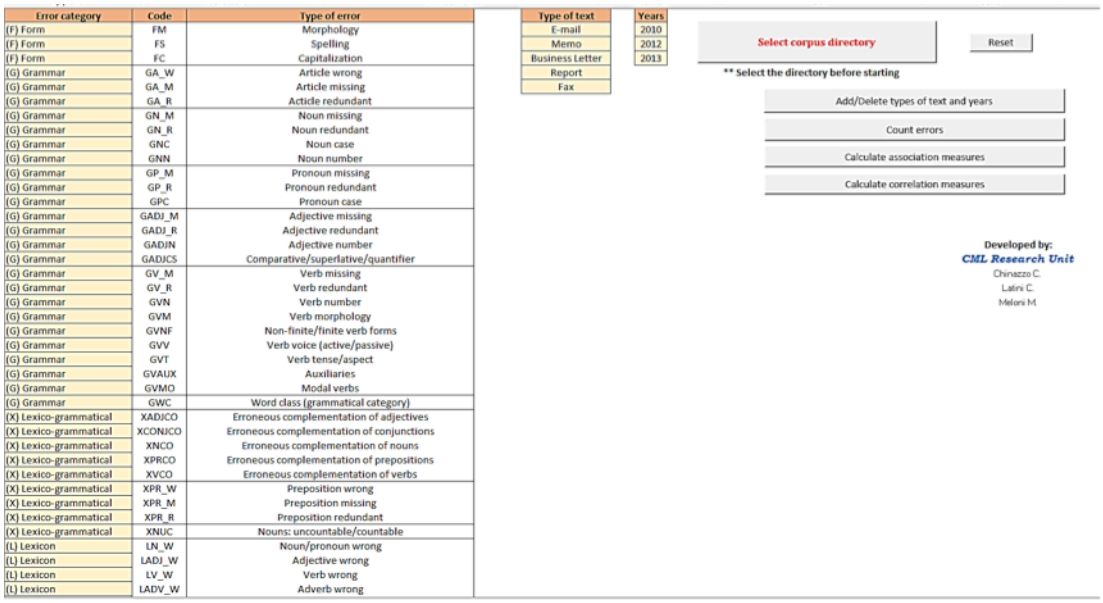
The VBA software has been specifically built to analyze a Corpus of 400 exam tests consisting in 152 business letters, 138 emails, 86 memos, 15 reports and 9 faxes written by Italian 2^{nd} level degree students studying Business English at the University of Tuscia in Viterbo, Italy. The VBA software can be used with any corpus in which the errors are identified in the sample texts by the abbreviations (inserted between round brackets) of the Louvain error tagging taxonomy as modified by Anna Romagnuolo for her Business Writing Corpus: e.g. a student's choice of the wrong genre type to complete an assigned written task is tagged with (TGM) = Textual genre mis-selection. The type of text (Letter, Memo, E-mail or Fax) and the year in which it was written are recorded in the file name in order to allow separate synchronic analysis of sample categories and diachronic analysis of learners' outputs per years (for example, XXX_E-mail_2012.txt). In the corpus, progressive numbers and student's name initials are associated to each sample text to allow retrieval of the original untagged version, which has also been saved in a scanned version usable with Access.

The software can determine the absolute and relative frequencies of specific errors (or error categories) and their degree of association by measuring both its significance by using chi-squared, log-likelihood, Z-score and T-score tests (they assess the acceptability of a starting hypothesis of non-association against the opposite hypothesis, with a probability of correctness of the results of 90%/95%/99%, and the possibility to make inferences on the entire population) ,

1. CML is an acronym for the software developers Cristiana **C**hinazzo, Claudio **L**atini and Mirko **M**eloni.

and its “size”, by applying relative measures such as Pearson’s correlation and the odds ratio. The graphical interface, which appears when running the software, is represented by an initial Excel sheet (figure 1) displaying three main elements:

- On the left, column 1 shows the error categories, with their tagging system derived from a modified Louvain taxonomy (column 2), explained in column 3;
- The small tables in the center summarize the types of text included in the corpus and the years in which they were collected;
- The options on the right allow the selection of the corpus, the addition or deletion of new texts and years to be analyzed and the type of measurements available.



Error category	Code	Type of error
(F) Form	FM	Morphology
(F) Form	FS	Spelling
(F) Form	FC	Capitalization
(G) Grammar	GA_W	Article wrong
(G) Grammar	GA_M	Article missing
(G) Grammar	GA_R	Acticle redundant
(G) Grammar	GN_M	Noun missing
(G) Grammar	GN_R	Noun redundant
(G) Grammar	GNC	Noun case
(G) Grammar	GNN	Noun number
(G) Grammar	GP_M	Pronoun missing
(G) Grammar	GP_R	Pronoun redundant
(G) Grammar	GPC	Pronoun case
(G) Grammar	GADJ_M	Adjective missing
(G) Grammar	GADJ_R	Adjective redundant
(G) Grammar	GADJN	Adjective number
(G) Grammar	GADJCS	Comparative/superlative/quantifier
(G) Grammar	GV_M	Verb missing
(G) Grammar	GV_R	Verb redundant
(G) Grammar	GVN	Verb number
(G) Grammar	GVM	Verb morphology
(G) Grammar	GVNF	Non-finite/finite verb forms
(G) Grammar	GVV	Verb voice (active/passive)
(G) Grammar	GVT	Verb tense/aspect
(G) Grammar	GVAUX	Auxiliaries
(G) Grammar	GVMO	Modal verbs
(G) Grammar	GWC	Word class (grammatical category)
(X) Lexico-grammatical	XADJCO	Erroneous complementation of adjectives
(X) Lexico-grammatical	XCONJCO	Erroneous complementation of conjunctions
(X) Lexico-grammatical	XNCO	Erroneous complementation of nouns
(X) Lexico-grammatical	XPRCO	Erroneous complementation of prepositions
(X) Lexico-grammatical	XVCO	Erroneous complementation of verbs
(X) Lexico-grammatical	XPR_W	Preposition wrong
(X) Lexico-grammatical	XPR_M	Preposition missing
(X) Lexico-grammatical	XPR_R	Preposition redundant
(X) Lexico-grammatical	XNUC	Nouns: uncountable/countable
(L) Lexicon	LN_W	Noun/pronoun wrong
(L) Lexicon	LADJ_W	Adjective wrong
(L) Lexicon	LV_W	Verb wrong
(L) Lexicon	LADV_W	Adverb wrong

Figure 1: Software graphic interface

In particular, the ”Count errors” option on the right allows a choice of/among texts to be analyzed (figure 2), with the possibility of creating a graph illustrating the overall results, as in figure 3, or partial results as in figure 4.

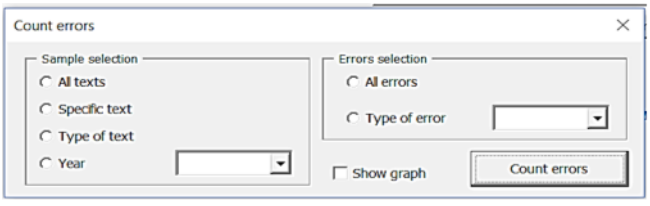


Figure 2: UserForm for error counting

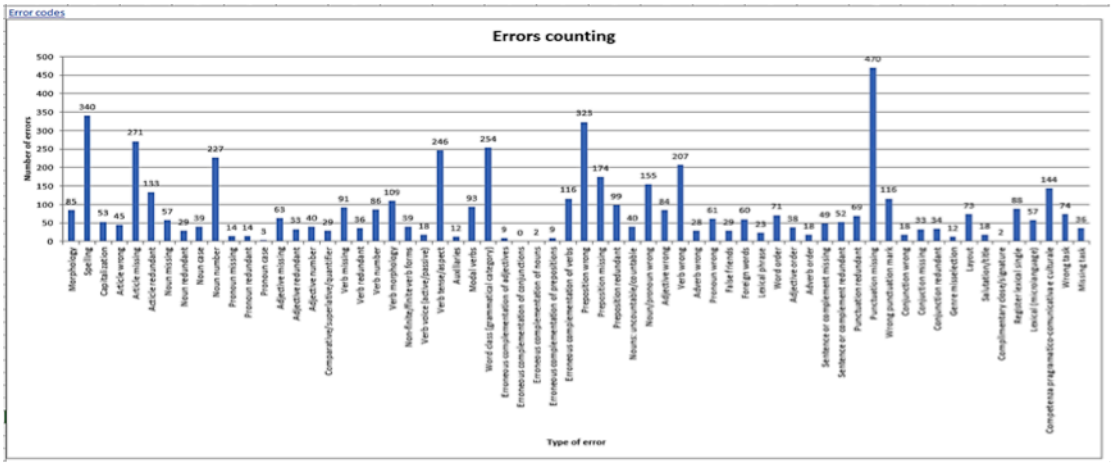


Figure 3: Error type frequency graph

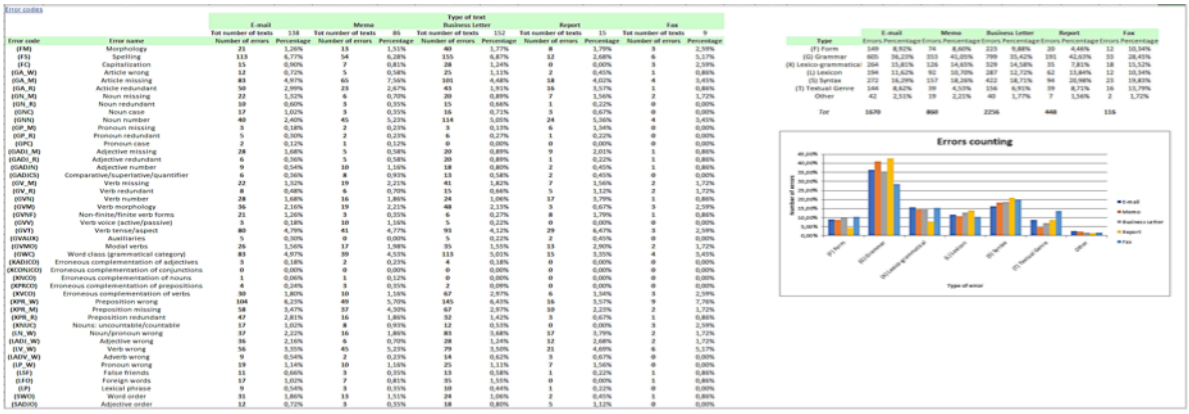


Figure 4: Error frequencies by category (on all errors)

A "*Sample selection*" function in the Count error window (figure 2) allows the user to choose to analyze all the texts, a specific text, all the texts related to a specific year or all the texts belonging to a specific category and a specific year. The "*Errors selection*" function allows the user to see all the errors or to select the ones belonging to a specific error type. By selecting the "*Show graph*" function, a graph will be created with a summary of the obtained results. Finally, by clicking on the "*Count errors*" button, the software will start its analysis. Once the count has started, it is possible to visualize a series of options (on the top right of figure 1) leading to the different visualizations of the results. The option "*Calculate association measures*", in the starting window (on the right of figure 1), opens a new dialogue box "*Association*" (figure 5) which allows the selection of the two variables to be analyzed (categories of errors or single errors). The "*Sample selection*" function offers the choice of the texts to analyze.

The analysis of the statistical association between two variables will establish whether the presence of one variable is linked to the simultaneous presence of the other (Borra & Di Ciaccio,

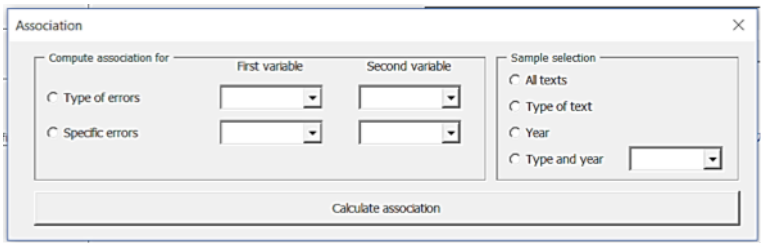


Figure 5: UserForm to compute association

2014). As in Hardi (2014), the measures computed are Chi-squared, Log-likelihood, T-Score, Z-Score, Odds Ratio.

In the opening window (figure 1), the selection of the option "*Calculate correlation measures*" will start the analysis of the correlation between variables (single errors or categories of errors) in the selected sample of texts. A first UserForm or dialogue box will be opened (figure 6) allowing users to choose between the analysis of "*All variables*" and "*Specific variables*".

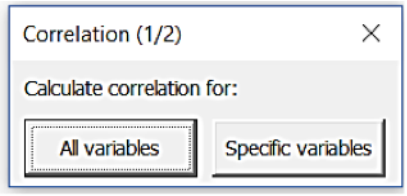


Figure 6: UserForm to compute correlation

Figure 7: UserForm to compute correlation

The selection of "*All variables*" will prompt the software to calculate the correlation matrix between all the errors. A matrix will be built to summarize the relationships between them (if these are chosen in the dialogue box shown in figure 7) or between all the categories of errors (if the option "*Type of errors*" is selected).

By clicking on "*Specific variables*" in the correlation dialogue box (figure 6), on the other hand, the software will only calculate the correlation between two specific variables (errors or categories of errors). A drop-down menu will open to offer the option of focusing on specific errors or specific text types.

The values of the correlation can be between -100% and +100%. A positive value means that the two variables are directly correlated: as the number of occurrences of the first type of error in a text increases, the second type of error is also expected to rise. If the value of the correlation is zero, the variables are not linearly correlated, so the dynamics of one do not influence the dynamics of the other. If the value is negative, the variables are inversely correlated: as the number of occurrences of the first type of error increases, the other is expected to decrease (Borra & Di Ciaccio, 2014). For both positive and negative values, however, the increase in absolute value of

the correlation indicates a closer link between the variables, whether direct or inverse. The closer the values are to zero, the weaker is the link between the two.

In order to test the accuracy of our software, the results of its basic function "Count errors" have been compared with the ones obtained by running the software package *WordSmith* tool (`https://www.lexically.net/wordsmith/`). As for the reliability of the remaining functions (associations and correlations), their results can be considered highly accurate since they depend on the approximation of the decimal numbers obtained by using Excel (which can have a maximum of 14 decimal digits). For example, the number 1.234567890123456 has 15 numbers after the decimal point and, as consequence, the last digit "6" is replaced by a 0 (1.234567890123450). This grade of approximation is retained to be sufficiently satisfactory.

References

Barone, L. (2010). "Computer Learner Corpora e sistemi di annotazione dell'errore". *Testi e linguaggi*, 4, pp. 121-139.

Borra, S., & Di Ciaccio, A. (2014). *Statistica: metodologie per le scienze economiche e sociali.* McGraw-Hill.

Cantos Gomez, P. (2011). *Statistical Methods in Language and Linguistic Research.* Lancaster: Equinox Publishing Limited.

Dagneaux, E., Denness, S., Granger, S., e Meunier, F. (1996). *Error Tagging Manual, Version 1.1, Center for English Corpus Linguistics.* Louvain-la-Neuve: Université Catholique de Louvain.

Dagneux, E., Denesse S., Granger, S. (1998) "Computer Aided Error Analysis". *System,* 2. 26, 163-174.

Ellis, R. e Barkhuizer, G. (Eds.). (2005). *Analysing Learner Language.* Oxford: OUP.

Granger, S., Hung, G., Petch-Tyson, S. (2002). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* Amsterdam: John Banjamins.

Hardie, A. (2014). *The statistics of collocation: from current practice to a new approach.* Invited talk at the Department of English, University of Uppsala, Sweden.

Index