CORPUS ORAUX, CORPUS ÉCRITS : PRATIQUES CROISÉES

Colloque co-organisé par le projet CORAL, les laboratoires Praxiling (UMR 5267), Dipralang (EA 739) et CRISES (EA 4424)

Lundi 13 novembre, de 9h30 à 12h30 et de 14h à 19h, Saint Charles 2, salle 003 « Caryatides »
Lundi 14 novembre, de 9h à 12h30 et de 14h à 18h, Saint Charles 2, salle 003 « Caryatides »

CORPUS ORAUX, CORPUS ÉCRITS : PRATIQUES CROISÉES

Conférenciers invités : Stefan Evert (FAU Erlangen-Nü rnberg), Bente Maegaard (Centre for Language Technologie, University of Copenhagen), Damon Mayaffre (BCL UMR 7320, Université Nice Sophia Antipolis), Christophe Parisse (Modyco UMR 7114, Université Paris Ouest Nanterre)

Argument

La constitution de ressources linguistiques représente depuis de nombreuses années l'une des productions majeures pour les chercheurs et les laboratoires en sciences humaines et sociales (SHS). Dans le domaine des sciences du langage, l'exploitation et l'analyse des données langagières ont connu un bond en avant avec l'apport des linguistiques de corpus et des traitements informatiques. D'une part, il s'avère indispensable de fédérer des pratiques scientifiques établies ou émergentes, relativement divergentes selon différents domaines (phonétique/phonologie, lexique, morpho- syntaxe, sémantique, linguistique des textes et des discours) et communautés. D'autre part, il s'agit, pour la linguistique outillée, de rendre accessibles ses acquis et méthodes à tous les chercheurs en SHS, qui fondent leurs travaux sur l'analyse de données langagières qu'elles relèvent de productions écrites, orales ou multimodales.

L'objectif de ce colloque est de faire dialoguer ces différentes communautés de chercheurs, qu'ils soient producteurs ou utilisateurs de ressources et/ou d'outils. Les approches méthodologiques qui associent différents modes d'exploitation et d'analyse de corpus oraux et écrits seront particulièrement attendues.

Il s'agira notamment d'aborder les problématiques suivantes :

- Annotation (phonétique, prosodique, gestuelle, morphosyntaxique, sémantique, etc.) ;
- Standardisation des annotations et du balisage ;
- Mise en place d'un format commun aux corpus écrits et oraux, permettant d'assurer l'interopérabilité entre les deux types de données ;
- Exploitation automatique ou semi-automatique et analyse des données ;
- Mutualisation des ressources (corpus annotés, schémas d'annotation, lexiques, patrons ou grammaires...) ;
- Adaptation et/ou création d'applications pour le traitement et l'exploitation des données orales et écrites.

Comité d'organisation : Beatrice Dal Bo (Praxiling, UM3), Sascha Diwersy (Praxiling, UM3), Christelle Dodane (Praxiling, UM3), Souad El Fellah (Praxiling, UM3), Hubert Gbedahou (Praxiling, UM3), Myriam Maréchal (UM3), Bénédicte Pivot (Dipralang, UM3), Bieke Van Camp (CRISES, EA 4424).

Comité scientifique :

- Nathalie Auger (Praxiling, UM3)
- Melissa Barkat-Defradas (Institut des Sciences de l'Evolution, UMR 5554, Montpellier 2)
- Sascha Diwersy (Praxiling, UM3)
- Christelle Dodane (Praxiling, UM3)
- Stefan Evert (FAU Erlangen-Nü rnberg)
- Francesca Frontini (Praxiling, UM3)

- Fabrice Hirsch (Praxiling, UM3)
- Agata Jackiewicz (Praxiling, UM3)
- Marie-Paule Jacques (Laboratoire LIDILEM, Université Grenoble Alpes)
- Dominique Legallois (Université Sorbonne Nouvelle, CLESTHIA)
- Giancarlo Luxardo (Praxiling, UM3)
- Bente Maegaard (Centre for Language Technologie, University of Copenhagen)
- Damon Mayaffre (BCL UMR 7320, Université Nice Sophia Antipolis)
- Rachel Panckhurst (Praxiling, UM3)
- Christophe Parisse (Modyco, Université Paris Ouest Nanterre)
- Bénédicte Pivot (Dipralang, UM3)
- Agnès Tutin (LIDILEM, Université Grenoble Alpes)

SPEECH CORPORA, TEXT CORPORA: SHARING PRACTICES

Invited speakers: Stefan Evert (FAU Erlangen-Nürnberg), Bente Maegaard (Centre for Language Technologie, University of Copenhagen), Damon Mayaffre (BCL UMR 7320, Université Nice Sophia Antipolis), Christophe Parisse (Modyco UMR 7114, Université Paris Ouest Nanterre).

Argument

The constitution of language resources represents one of the major productions for researchers and laboratories in social sciences and humanities (SSH) in many years. In the field of language sciences, the exploitation and analysis of linguistic data have made a significant leap forward with the contribution of corpus linguistics and computer processes. Firstly, it proves vital in uniting established and emerging scientific practices, relatively divergent according to different fields (phonetics/phonology, lexicon, morphosyntax, semantics, text linguistics and discourse linguistics) and communities. Secondly, it is a matter of making the benefits and methods of tool-assisted linguistic analysis accessible to all SSH researchers, who base their works on the analysis of linguistic data, whether text, speech or multimodal productions.

The goal of this symposium is to create a dialogue between these different communities of researchers, whether they be producers or users of resources and/or tools. The methodological approaches that associate different modes of exploitation and analysis of speech and text corpora will be particularly welcomed.

The main focus will be on addressing the following issues:

- Annotation (phonetic, prosodic, gestural, morphosyntactic, semantic, etc.);
- Standardization of annotations and markups ;
- Establishment of a common format for speech and text corpora, allowing for the assurance of interoperability between the two types of data ;
- Automatic or semi-automatic exploitation and data analysis;
- Sharing of resources (annotated corpora, annotation schemas, lexicons, patterns or grammars...);
- Adaptation and/or creation of applications for the processing and exploitation of speech and text data.

Organizing committee: Beatrice Dal Bo (Praxiling, UM3), Sascha Diwersy (Praxiling, UM3), Christelle Dodane (Praxiling, UM3), Hubert Gbedahou (Praxiling, UM3), Myriam Maréchal (UM3), Bénédicte Pivot (Dipralang, UM3), Bieke Van Camp (CRISES, EA 4424)

Programme committee:

- Nathalie Auger (Praxiling, UM3)
- Melissa Barkat-Defradas (Institut des Sciences de l'Evolution, UMR 5554, Montpellier 2)
- Sascha Diwersy (Praxiling, UM3)
- Christelle Dodane (Praxiling, UM3)
- Stefan Evert (FAU Erlangen-Nürnberg)
- Francesca Frontini (Praxiling, UM3)
- Fabrice Hirsch (Praxiling, UM3)
- Agata Jackiewicz (Praxiling, UM3)
- Marie-Paule Jacques (Laboratoire LIDILEM, Université Grenoble Alpes)

- Dominique Legallois (Université Sorbonne Nouvelle, CLESTHIA)
- Giancarlo Luxardo (Praxiling, UM3)
- Bente Maegaard (Centre for Language Technologie, University of Copenhagen)
- Damon Mayaffre (BCL UMR 7320, Université Nice Sophia Antipolis)
- Rachel Panckhurst (Praxiling, UM3)
- Christophe Parisse (Modyco, Université Paris Ouest Nanterre)
- Bénédicte Pivot (Dipralang, UM3)
- Agnès Tutin (LIDILEM, Université Grenoble Alpes)

Sommaire des résumés

VariaIdade Corpus oraux, corpus écrits

Ronny Beckert (Romanisches Seminar Universität Heidelberg, Allemagne)

Le projet de recherche « Varia-Idade no Rio de Janeiro – Comunicação e geração : Estratégias linguísticas e discursivas na idade maior » (*Varia-Idade à Rio de Janeiro – Communication et génération : Des stratégies linguistiques et discursives des personnes âgées*) est une coopération entre l'Université de Heidelberg et l'Université d'État de Rio de Janeiro. Dans le cadre de ce projet sera créé un corpus d'environ 100 entretiens traitant de la vie et des expériences individuelles dans la ville. Les participants de cette étude ont au moins 60 ans et résident dans différents quartiers de la ville de Rio de Janeiro depuis plus de 40 ans.

Outre l'analyse du comportement linguistique et de la variation linguistique dans la parole des habitants plus âgés de Rio de Janeiro, sera analysé le discours sur la perception des changements dans le quotidien et dans l'espace urbain. Rio de Janeiro a souffert et continue à souffrir un changement global pendant les dernières 50 années. Le changement dans l'espace urbain est souvent une question des *fractures urbaines* qui se reflètent également au niveau métalinguistique.

Le comportement linguistique de la génération des jeunes a été bien décrit et analysé pendant les dernières décennies, alors que le comportement linguistique des personnes âgées était rarement au centre de l'attention. Beaucoup d'études existantes traitant du comportement linguistique des personnes âgées partent de l'hypothèse que la communication des locuteurs en âge avancé soit déficitaire.

Pour l'analyse du discours urbain des locuteurs plus âgés il est impératif de s'approcher au sujet d'une perspective transdisciplinaire qui établit un lien de la perspective linguistique et analytique- discursive avec la perspective de la géographie urbaine, de la sociologie, de l'histoire, de l'anthropologie urbaine et de la science culturelle.

Ma proposition pour votre colloque est de présenter le projet ainsi que le corpus, de préférence, à travers une communication orale.

Bibliographie :

Gerstenberg, Annette (2011): *Generation und Sprachprofile im höheren Lebensalter. Untersuchungen zum Französischen auf der Basis eines Korpus biographischer Interviews*, Frankfurt am Main: Klostermann.

Grosse, Sybille / Tsekos, Nicolas / Bulot, Thierry (1996) : « L'évaluation en discours : la mise en mots des fractures urbaines », dans : Richard-Zapella, Jeannine : *Le questionnement social : Actes du colloque international de Rouen 16-17 mars 1995*, Paris 1996, 295-302.

Grosse, Sybille (1999) : « Vitalité linguistique et dynamique langagière : le berlinois », dans : Thierry Bulot / Nicolas Tsekos (éds.) : *Langue urbaine et identité : langue et urbanisation linguistique à Rouen, Venise, Berlin, Athènes et Mons*, Paris ; Montréal : Harmattan, 127- 151.

Preti, Dino (1991) : *Linguagem dos Idosos : Um estudo de análise da coversação,* São Paulo : Contexto.

Enjeux méthodologiques et technologiques pour la constitution d'un corpus d'arabe tunisien

Yossra Ben Ahmed, Fatma Ben Barka et Linda Hriba (Laboratoire Ligérien de Linguistique (LLL), Université d'Orléans)

La constitution et la mise à disposition de corpus de parlers arabes se heurtent à l'absence de ressources et au manque d'outils pour le traitement de ces derniers. Aussi, lors de la construction d'un corpus de l'une des variétés, i.e. l'arabe tunisien, nous avons été confrontées à de maints problèmes pour lesquels nous avons dû opérer des choix méthodologiques que nous décrivons dans cette communication. L'élaboration et l'exploitation d'un corpus de 30 heures d'enregistrement de l'arabe tunisien réalisé dans le cadre d'études doctorales (en cours) pour l'analyse de l'usage du futur et du subjonctif, a nécessité la mise en oeuvre d'un ensemble de procédures, allant du recueil de données jusqu'aux étapes d'annotations.

Aussi dans un souci de comparabilité, nous nous sommes d'abord appuyées sur la méthodologie utilisée dans le cadre du projet ESLO, un grand corpus de données orales, constitué de trois enquêtes :

- ESLO1 (1968-1970), corpus clos de 470 enregistrements (4,5 millions de mots)
- ESLO2, (depuis 2000), le but est d'atteindre plus de six millions de mots (450 heures d'enregistrements)
- ESLO-LCO (2008-2014) - l'objectif est l'étude de la vie des langues en contact avec le français.

  Le corpus d'arabe tunisien a été essentiellement enregistré à Orléans[1] (et agglomération), pour l'entretien en face à face nous avons utilisé la trame d'enquête élaborée par les membres du projet ESLO et pour satisfaire aux conditions d'une recherche en sociolinguistique l'accent a été mis, sur la diversité des situations (entretiens, repas, interviews de personnalités) et sur des profils sociologiques variés.

  Une fois les 30 heures d'enregistrements réalisés et afin de permettre l'analyse et le traitement des données, nous avons dû entamer les phases de transcription et d'annotation, deux phases complexes pour lesquelles il manque un cadrage théorique et méthodologique pour l'arabe tunisien.

En effet, les travaux sur l'arabe tunisien, peu nombreux, hésitent entre les deux systèmes graphiques, i.e. latin et arabe. Le choix de l'un ou l'autre système est dicté par de nombreux paramètres, allant de la tradition du champ, jusqu'aux préférences idéologiques, en passant par la facilité technique. C'est pour cette dernière raison, que nous avons opté pour une transcription avec une graphie latine. Quant au mode de transcription, nous avons opté pour une notation orthographique, même si ce choix n'est pas, ainsi que nous tenterons de le montrer, sans poser de nombreux problèmes en l'absence d'un standard stabilisé. Nous exposerons les conventions choisies en les comparant aux principales conventions proposées. En ce qui concerne l'outil de transcription, nous avons choisi TRANSCRIBER[2]. Nous tenterons de montrer,

[1] Les premiers enregistrements ont été réalisé à Orléans (37 entretiens) mais rapidement nous avons dû réaliser quelques enregistrements en Tunisie, enregistrer des personnalités et des repas s'avérant plus aisé.

[2] Téléchargeable sur : http://www.ldc.upenn.edu/mirror.Transcriber/

les avantages qu'offre cet outil. Autre point, l'absence d'un étiqueteur morpho-syntaxique, pour exploiter le corpus constitué, nous a conduit à baliser manuellement les occurrences pour le futur et le subjonctif dans un fichier Transcriber. Ces derniers ainsi identifiées ont été extraites grâce au logiciel d'analyse textométrique TXM[3], et exportées dans un tableau CSV, afin d'y être annotées. Chacune des occurrences a ainsi été sous-spécifiée pour un certain nombre de traits morphosyntaxiques et sémantiques, dont nous présenterons brièvement les structurations arborescentes. L'objectif de cette communication est donc double : (i) après avoir exposé les différents éléments méthodologiques du projet ESLO réinvestis pour l'élaboration de notre corpus de l'arabe tunisien,
(ii) nous reviendrons sur les points problématiques pour lesquels nous avons dû opérer des choix afin de rendre le corpus disponible et interopérable.

Bibliographie :

Abouda L. (2015). Syntaxe et Sémantique en corpus. Du temps et de la modalité en français oral, mémoire HDR, Université d'Orléans.

Abouda, L. & Skrovec, M. (2015). « Du rapport entre formes synthétique et analytique du futur. Étude de la variable modale dans un corpus oral micro-diachronique », *Revue de Sémantique et Pragmatique*, 38, 35-57.

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

Baude, O. (2006). *Corpus oraux, Guide des bonnes pratiques*. Paris, CNRS Editions Benjelloun S. (2002). Une double graphie, latine et arabe, pour enseigner l'arabe marocain,

in: D. Caubet, S. Chaker, J. Sibille (éds), Codification des langues de France, 331-340, L'Harmattan, Paris.

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

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

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

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

[3] Téléchargeable à l'adresse suivante : http://textometrie.ens-lyon.fr/

Ce que disent des élèves sur les classes coopératives en collège et lycée

Sylvain Connac (Université Paul Valéry de Montpellier, LIRDEF)

Des classes coopératives voient le jour dans des collèges et lycées de l'enseignement public. Cette recherche phénoménologique (Merleau-Ponty, 1945) s'intéresse à ce que pensent les élèves inscrits dans ces classes. Elle propose une synthèse d'un corpus de 22 entretiens semi- directifs, réalisés au sein de 11 classes de deux collèges et d'un lycée. Trois établissements (11 classes) ont servi de support: un collège en éducation prioritaire (12 entretiens d'élèves en 6ème), un collège en milieu rural (1 élève en 6ème et 4 élèves en 5ème) et un lycée d'enseignement général et technologique (2 élèves en seconde).

L'intérêt d'écouter ce qu'expriment des élèves de classes coopératives est multiple. En effet, ce sont eux qui vivent au quotidien ce que les enseignants parviennent à organiser en matière de pratiques coopératives. Ils sont donc à même de témoigner de la nature et des effets de ces approches pédagogiques. D'autant plus, qu'ne tant que collégiens ou lycées, leur scolarité passée les a conduits à rencontrer d'autres façons d'enseigner. Ils se présentent ainsi tels des révélateurs, apportant une distanciation quant à ce que des enseignants pourraient dire de ces mêmes réalités. N'étant pas à l'origine de ces projets et ne pouvant être accusés de quelconque idéologie, ils présentent un vécu ordinaire, tel que compris personnellement. De plus, écouter des élèves s'exprimer sur la façon dont ils travaillent permet d'entrer dans le grain fin du fonctionnement d'une classe coopérative. Ils proposent une visite détaillée, à partir de ce qu'ils estiment important et qui peut différer de ce que souligneraient des adultes. Cette recherche du détail des organisations coopératives est permise par le croisement des propos collectés par les entretiens. Chaque élève ne dit pas tout, mais la complémentarité de leurs discours présente une pluralité de fonctionnements de classe. Enfin, ce que disent les élèves est à l'image de leur posture en classe, on accède ainsi à une partie de ce que forment les projets de classes coopératives en matière de « métiers d'élèves » (Perrenoud, 2004), entendus ici comme des prédispositions mentales dans le rapport à la scolarisation.
Les entretiens ont eu une durée moyenne de 30 minutes chacun. Chacun a été enregistré, puis entièrement retranscrit. Pour traiter les données, nous avons réalisé des analyses de contenus catégorielles[4] (Bardin, 1997) Elle est définie comme une « méthode de classification ou de codification dans diverses catégories des éléments du document analysé pour mieux en faire ressortir les différentes caractéristiques en vue d'en mieux comprendre le sens exact et précis.
» (L'écuyer, *in* Deslauriers, 1988, p. 50) Nous n'avons pas déterminé d'hypothèse exploratoire préalable. Aussi, les conclusions de ce travail prennent la forme d'une synthèse de la palette des avis collectés.
Une fois chaque entretien retranscrit et analysé, nous avons réalisé cette synthèse en croisant les idées énoncées individuellement, sans n'en rejeter aucune. La prise en compte des derniers entretiens a révélé de nombreuses redites, signes d'une globale exhaustivité des idées présentées.

Bibliographie

Bardin, L. (1997). *L'analyse de contenu*. Paris : PUF.
Connac, S. (2016). Autonomie, responsabilité et coopération : ce qu'en disent les élèves utilisant un plan de travail, *Éducation et socialisation* [En ligne], *41 | 2016*, URL : http://edso.revues.org/1725
Connac, S. (2017 a). *La coopération entre élèves*. Futuroscope : Editions Canopé.
Deslauriers, J.P. et al. (1988). *Les méthodes de la recherche qualitative*. Sillery : Presses de l'Université du Québec.
Merleau-Ponty, M. (1945). *Phénoménologie de la perception*. Paris : Gallimard.
Perrenoud, P. (2004). *Métier d'élève et sens du travail scolaire*. Issy-les-Moulineaux : ESF Editeur.

[4] L'analyse de contenu catégorielle expliquée à des étudiants en Sciences de l'Education https://www.youtube.com/watch?v=MNWq9-IkYvA

Un corpus écrit et oral annoté et des outils d'interrogation spécifiquement élaborés pour faciliter des études grammaticales sur le français contemporain

Jeanne-Marie Debaisieux, José Deulofeu, Frédéric Béchet, Alexis Nasr (Laboratoire d'Informatique Fondamentale, Aix-en-Provence)

La plate-forme ORFEO offre des ressources sur le français contemporain et des outils pour les exploiter dans le cadre de la recherche en linguistique et en TAL ou pour l'enseignement de la langue. Le Corpus d'Etudes sur le Français Contemporain rassemble plus de 10 M. de mots à l'écrit et à l'oral, échantillonnées selon les standards internationaux : transcriptions d'enregistrements récents de plus de 2000 locuteurs différents de toutes les régions de France ainsi que de la Suisse et de la Belgique et textes extraits de littérature, de presse quotidienne ou régionale, de textes scientifiques et d'écrits non professionnels (SMS, Blog). Les transcriptions des enregistrements et le son sont alignés au phonème. Plusieurs formats de son sont proposés selon la qualité et le temps de téléchargement. Les textes écrits et les transcriptions ont été annotés semi-automatiquement selon le lemme, la catégorie (nom, verbe, adjectif, etc.) ou la fonction syntaxique. L'originalité de ces annotations tient au souci d'en faire un outil robuste pour la recherche de structures grammaticales. Des tags nouveaux (PERIPH, DM, PARENTH, PARA) ont été définis pour prendre en compte les constructions fréquentes dans la langue parlée (détachement, marqueurs de discours, parenthétiques, effets de liste et amorces). D'autres annotations ont été volontairement sous-spécifiées (tag par défaut DEP). L'objectif est en effet de donner le maximum de fiabilité au travail de recueil des données au moyen de l'outil d'extraction d'arbres : il s'agit de privilégier le rappel plutôt que la précision afin d'éviter la perte de données cruciales pour l'argumentation linguistique. Dans ce même esprit, un soin particulier a été apporté au traitement des séquences potentiellement ambiguësentre une lecture libre et une lecture en expression fonctionnelle multimots : « je sais bien que tu es parti » versus « je le sais bien que tu ne m'aies rien dit » (Nasr & coll 2014). La structure des syntagmes à tête non verbale à lecture partitive (*de l'eau, beaucoup d'eau*) a reçu un traitement unifié avec pour tête l'élément le plus à gauche (Ramisch 2016). Enfin, le dictionnaire a été construit à partir du LEFF en fonction des avancées de la linguistique descriptive : catégorie DET définie de façon restrictive différenciée de la relation SPE, regroupement des relatifs-interrogatifs dans une catégorie unique de « mots Q ». L'annotation automatique a été réalisée grâce à l'entrainement sur les données annotées à la main avec l'outil Macaon (Nasr 2011) pour l'oral et FRMG pour l'écrit (Villemonte de la Clergerie 2013). L'ensemble des données est en accès libre et téléchargeable. La recherche peut se faire en fonction du type de texte (écrit ou oral, presse ou littérature) mais aussi de critères de provenance géographique, d'âge ou de nombre d'intervenants dans les enregistrements transcrits. Des exemples de résultats de requêtes (relatives versus consécutives) seront donnés et comparés à une recherche manuelle sur un sous- corpus.

Bibliographie

De la Clergerie, E., Sagot, B., Nicolas, L., Guenot, M.-L., FRMG : évolutions d'un analyseur syntaxique TAG du Français, ATALA.

Nasr A., Bechet, F., Favre, B., Bazillon, T., Deulofeu, J., Valli, A. 2014. *Automatically Enriching Spoken Corpora with Syntatic Information for Linguistic Studies* International Conference on language Resources and Evaluation (LREC).

Nasr A., Bechet, F., Rey, J.-F., *MACAON Une chaîne linguistique pour le traitement des graphes de mots.* Traitement Automatique des Langues Naturelles.

Ramisch C., Narsr A., Valli A., Deulofeu J., (2016) DeQue : a lexicon of complex prepositions and conjunctions in French, Proceedings of the 10th LREC conference.

Parole disfluentes : un corpus de parole bègue

Ivana Didirkova (Université catholique de Louvain), **Fabrice Hirsch** et **Giancarlo Luxardo** (Praxiling, Montpellier 3)

Le bégaiement est un trouble de l'élocution qui touche 1% de la population mondiale. Il se caractérise notamment par des altérations sévères de la fluence intervenant le plus généralement durant un acte de communication.

Si plusieurs travaux ont été et sont menés sur le sujet en sciences du langage, les chercheurs se heurtent souvent à la difficulté d'entrer en contact avec des personnes souffrant d'un bégaiement. Cette difficulté peut s'expliquer par sa faible prévalence mais aussi par le fait que les personnes qui bégaient peuvent se montrer hésitantes à l'idée que leur parole puisse être enregistrée, écoutée et analysée. C'est la raison pour laquelle plusieurs chercheurs ont entrepris de mettre à la disposition de la communauté scientifique les enregistrements qu'ils ont acquis lors de différentes recherches menées sur le bégaiement. Ainsi, divers projets, tels que la *FluencyBank* (http://fluency.talkbank.org/) ont émergé ces dernières années et proposent des enregistrements de personnes qui bégaient en anglais et en allemand.

Le corpus *Paroles disfluentes* se veut être un prolongement à ces différentes bases de données en proposant des enregistrements en français. Il est constitué de fichiers audio obtenus dans le cadre de différentes recherches réalisées au laboratoire Praxiling de Montpellier (par ex. Goyet *et al.*, 2013 ; Didirkova *et al.*, 2015 ; Didirkova, 2016). Plus précisément, *Paroles disfluentes* est constitué de 38 enregistrements de personnes Adultes Qui Bégaient (AQB) se prêtant à différents exercices. Les locuteurs sont au nombre de 17 (4 femmes et 13 hommes) et leur moyenne d'âge est de 32 ans (avec un écart-type de 11 ans). La durée totale du corpus est de 116 minutes avec des fichiers (au format *.wav*) allant de 45 secondes à 16 minutes 22 secondes.

Provenant de diverses études, le corpus contient différentes situations élocutoires produites par des personnes qui bégaient. On dénombre ainsi 10 fichiers de lecture dans lesquels les locuteurs lisent des textes, comme *La Chèvre de Monsieur Seguin* ou la fable *Le Lion et le rat*, et 12 enregistrements audio de parole spontanée. Concernant cette tâche, les AQB devaient raconter leur journée type ou le dernier livre lu. En outre, l'ensemble des enregistrements issus d'une étude sur les effets de double tâche sur le bégaiement figurent dans le corpus. Pour ces locuteurs, il s'agissait de résumer à l'oral des contes pour enfant, d'abord en tâche de parole simple, puis en effectuant un calcul mental ou encore en jouant à un jeu de pingpong sur ordinateur. Les enregistrements présents dans le corpus sont systématiquement accompagnés de transcriptions orthographiques sous différents formats (*TextGrid*, *.xls, .docx*). Celles-ci ont parfois donné lieu à un questionnement. En effet, le décompte des répétitions, la durée des prolongations et des blocages constituent des traits acoustiques propres aux disfluences qui pourraient être utiles aux chercheurs. Une réflexion a donc été engagée sur la transcription de ces disfluences et sur la manière de conserver leurs caractéristiques.

L'ensemble des fichiers audio et des transcriptions dont il est question dans ce résumé figurent sur la plateforme Ortolang. Le colloque *Corpus oraux, corpus écrits : pratiques croisées* sera l'occasion de présenter *Paroles disfluentes* et de discuter des problèmes méthodologiques liés à la transcription de personnes qui bégaient.

Bibliographie

Goyet L., Hirsch F. (2013) Etude des conséquences des situations de double tâche chez les personnes qui bégaient. Vèmes Journées d'Etudes sur la parole

Didirkova I., Fauth, C., Hirsch, F., Luxardo, G., Diwersy, S. (2016) Disfluences normales vs. Disfluences sévères : une étude acoustique. Actes de la conférence conjointe JEP-TALN-RECITAL 2016, vol.1 : JEP, 191-199, communication affichée

Didirkova I. (2016) Parole, langues et disfluences : une étude linguistique et phonétique du bégaiement. Thèse de Doctorat soutenue à l'Université Paul Valéry-Montpellier 3, 412 p.

Transcription d'un corpus de récits de vie de migrants en vue de son annotation sémantique en lieux et sentiments

Catherine Dominguès et Carmen Brando (COGIT, LaSTIG, IGN, Vincennes)

Les témoignages directs constituent une modalité de la compréhension de certains événements historiques. Dans ce contexte, le projet pluridisciplinaire Matriciel [1] vise à analyser (sous l'angle des lieux) pour les cartographier un corpus oral de récits de vie de républicains espagnols ayant migré en France en 1936. Le projet revêt ainsi un aspect patrimonial par la conservation et la mise en valeur de la parole des migrants, un corpus sur cette thématique n'existant pas à notre connaissance. Le travail vise l'identification automatique des noms de lieux et des sentiments associés, pour la représentation cartographique [2]. La notion de lieu recouvre noms propres (*Barcelone*, *Barcelona*) et noms communs (*camp de Gurs*, *camp de concentration*).

Afin de procéder à l'annotation sémantique avec des outils de TAL, trois états différents ont été élaborés :

- le corpus oral où la parole est qualifiée d'atypique parce que les locuteurs présentent des accents et des séquences en espagnol, allemand et arabe figurent dans les entretiens réalisés en français
- une transcription détaillée obtenue en adaptant les conventions ESLO [3] et ESTER2 [4], dans laquelle tours de parole, mots étrangers et traduction, contexte sonore, pauses courtes et longues, sont transcrits. Les disfluences (répétitions, révisions), les marqueurs de l'oral sont annotés. Outre les caractéristiques de l'accent espagnol [5], les entretiens contiennent des mots espagnols dont la prononciation est "francisée" comme *multe*, annoté *amende [pronFran=multe]* où *multa* en espagnol signifie amende, et des mots français dont la prononciation est "hispanisée" comme: *ferma annoté ferme [pronHisp= ferma]* où *ferme* se dit *granja* en espagnol
- une transcription simplifiée : la transcription détaillée ne permettant pas l'utilisation d'outils de TAL une transcription simplifiée est établie dans laquelle disfluences, marqueurs de discours, séquences en langue étrangère ont été supprimés ; seules les traductions sont conservées ; les pauses longues sont transformées en points.

Sur cette transcription simplifiée les lieux, essentiels pour la cartographie, sont annotés dans GATE [5] (qui enchaîne des outils de TAL construits ad hoc ou existants comme recherche d'information, annotation, POS) : les noms propres sous forme française ou espagnole, les noms communs en utilisant des techniques d'apprentissage à partir d'un corpus balisé manuellement.

Les mots porteurs de sentiments (polarités négative, neutre, positive) ont été repérés à l'aide d'un lexique de sentiments (adapté du lexique d'EMOTAIX [6]). Des phénomènes linguistiques, comme la négation, ont été identifiés grâce à l'adaptation au français des travaux existants [7]. L'objectif est maintenant de propager les polarités aux noms de lieux.

Ce travail est en cours. La discussion pourra porter sur :

- la possibilité d'aligner corpus oral et corpus écrit à l'aide d'annotations en couches comme décrit dans TCOF [8] et [9]
- la caractérisation quantitative et qualitative d'un corpus à partir de métadonnées spécifiques de l'oral (disfluences, alternance des langues, accents, etc.) sur la base d'un encodage TEI [10]
- la mise à disposition, par exemple sur ORTOLANG [11], pour les associations de migrants et les communautés scientifiques, d'une version anonymisée du corpus qui privilégie une approche comme [12].

Bibliographie :

[1] Dominguès C., Weber S., Brando C., Jolivet L., Van Damme M.-D. (2017, à paraître). "Analyse et cartographie des sentiments dans des récits de vie de migrants", Colloque international de géomatique et d'analyse, SAGEO'2017, Rouen

[2] Caquard S., Cartwright W. (2014). "Narrative cartography: From mapping stories to the narrative of maps and mapping". *The Cartographic Journal*, vol. 51, n°2, p. 101-106, http://dx.doi.org/10.1179/0008704114Z.000000000130

[3] ESLO, http://eslo.huma-num.fr/images/eslo/pdf/GUIDE_TRANSCRIPTEUR_V4_mai2013.pdf

[4] ESTER2, http://www.afcp-parole.org/camp_eval_systemes_transcription/docs/Conventions_Transcription_ESTE R2_v01.pdf

[5] Boula de Mareüil P., Vieru-Dimulescu B., Woehrling C., Adda-Decker M. (2008). "Accents étrangers et régionaux en français", *Traitement Automatique des Langues* 2008 Volume 49 Numéro 3 p.135-163

[5] GATE, https://gate.ac.uk/

[6] EMOTAIX, http://centrepsycle-amu.fr/outils-recherche/

[7] Andreevskaia A., Bergler S. (2007). "CLaC and CLaC-NB: Knowledge-based and corpus- based approaches to sentiment tagging", In: Proceedings of SemEval-2007: 4th International Workshop on Semantic Evaluations at ACL 2007, Prague http://www.aclweb.org/anthology/S/S07/S07-1022.pdf

[8] André V., Canut E. (2010). "Mise à disposition de corpus oraux interactifs : le projet TCOF (Traitement de Corpus Oraux en Français)", *Pratiques* [En ligne], 147-148 | 2010, mis en ligne le 15 décembre 2010, consulté le 25 juillet 2017. URL : http://pratiques.revues.org/1597 ; DOI : 10.4000/pratiques.1597

[9] Christodoulides G., Barreca G. (2017). "Expériences sur l'analyse morphosyntaxique des corpus oraux avec l'annotateur multi-niveaux DisMo", *Corela* [En ligne], HS-21 | 2017, mis en ligne le 16 février 2017, consulté le 21 février 2017. URL : http://corela.revues.org/4867

[10] http://www.tei-c.org/index.xml

[11] https://www.ortolang.fr/

[12] Panckhurst R. (2016). "A digital corpus resource of authentic anonymized French text messages: 88milSMS-What about transcoding and linguistic annotation?", Digital Scholarship in the Humanities. Published by Oxford University Press on behalf of EADH. http://dx.doi.org/10.1093/llc/fqw049, 11 pages.

Corpus oraux et corpus écrits en analyse du discours : quels outils/transcriptions pour l'observable langagier ?

Souad El Fellah (Praxiling, Montpellier 3)

Les récentes recherches en analyse du discours (AD) et en analyse conversationnelle (AC) tendent vers l'étude de l'observable langagier en tenant compte de tous les paramètres de son actualisation. D'emblée, les phénomènes langagiers sont pris en considération dans un corpus qu'il soit oral ou écrit. Ce matériau de base est re-présenté sous différentes formes selon le besoin de la recherche et l'objectif du chercheur. Ces différentes re-présentation résultent du recours aux outils de traitement de texte qu'ils soient automatiques ou semi-automatiques. Néanmoins, la transcription du matériau de base moyennant tels ou tels logiciels/transcriptions pose le problème du choix de l'outil et de l'annotation « pertinents » pour appréhender l'observable langagier dans son co(n)texte de production. Autrement dit, l'étude de l'observable langagier ne se limite pas à sa place et à sa fonction dans l'énoncé, elle prend en considération le cotexte ainsi que les paramètres extralinguistiques d'où la nécessité du choix de l'annotation (phonétique, prosodique, gestuelle, morphosyntaxique, sémantique, etc.) adéquate

A la lumière d'un projet de recherche personnel (L'apostrophe dans les *Questions au gouvernement*), ma contribution dans cette manifestation scientifique abordera la problématique du choix de l'annotation et son impact sur l'objet de recherche et l'objectif à atteindre. Notre propos démontrera le rôle du corpus oral et/ou écrit transcrits via des outils spécifiques dans l'étude de l'observable langagier. Aussi se projettera-t-il sur l'importance de réfléchir sur la mise en place pour les sciences du langage d'annotations/transcriptions *communes aux corpus écrits et oraux, permettant d'assurer l'interopérabilité entre les deux types de données*.

Bibliographie :

- Cori Marcel & Sophie David, 2008, « Les corpus fondent-ils une nouvelle linguistique ? », Langages
- Maingueneau D. 1987, *Nouvelles tendances en analyse du discours*. Langue Linguistique Communication. Hachette. Paris
- Valette Mathieu, 2008, *Textes, documents numériques, corpus pour une science des textes instrumentée*
- Bilger Mireille, 2000, *Méthodologie et applications linguistiques*
- Convention de transcription : ICOR
- ELAN/ CLAN

Expériences et recommandations pour la gestion de corpus de parole pathologique

Alain Ghio, Gilles Pouchoulin et Antoine Giovanni (LPL, Aix-en-Provence)

Actuellement, l'étude des dysfonctionnements de la voix et de la parole est sortie du simple cadre de la recherche clinique et intéresse les laboratoires de recherche issus des sciences du langage ou du traitement automatique de la parole. Par l'observation des dysfonctionnements, les chercheurs non-cliniciens confrontent les résultats de leur recherche établis sur des corpus de parole "normale" à des situations de dysfonctionnement. Le défi est immense car le cadre "pathologique" induit une variation considérable dans ses manifestations de surface, c'est-à-dire sur les productions sonores. Aux symptômes de la maladie se superposent les effets variables des traitements mais aussi des phénomènes de compensation non-uniformes des locuteurs. De ce fait, toute généralisation à une population clinique particulière nécessite l'observation d'un grand nombre de patients du fait de la très forte variation interindividuelle rencontrée. Il est donc important de mutualiser les enregistrements existants. Or pour être utilisables, ces enregistrements doivent répondre à de fortes exigences. Si les problèmes de prise de son ou autres signaux physiologiques sont en passe de devenir anecdotiques grâce à la diffusion de matériels de qualité et à la meilleure formation des personnels en charge des enregistrements, si le stockage des signaux de parole ne constitue plus actuellement un obstacle, si le recours à du matériau linguistique suffisant se généralise, le maillon faible reste la normalisation et la structuration des données sur les locuteurs et leurs productions langagières. Concrètement, si les données sonores sont souvent accessibles, elles ne présentent au final aucun intérêt si les liens entre les enregistrements et les caractéristiques cliniques du locuteur sont rompus ou erronés. L'objectif de ce travail est de présenter différentes actions de terrain et de proposer des recommandations pour la structuration des données sonores, physiologiques et cliniques (Ghio et al., 2012).

La base de données SPEEDI-DB (Speech Disorders Database) propose un cadre opérationnel pour stocker, archiver et partager des données de parole pathologique. Contrairement aux bases de données orales de type patrimonial, dialogal, conversationnel, la parole pathologique nécessite la collecte et la mémorisation précise des informations sur les locuteurs et le contexte d'enregistrement. Par conséquent, nous préconisons de renseigner au maximum les informations sociolinguistiques mais aussi médicales, symptomatiques (ex : date d'apparition et localisation des symptômes), contextuelles (ex : « le patient a pris ses médicaments 4h auparavant »). De même, toute forme d'évaluation (ex : UPDRS pour les malades de Parkinson, GRBAS pour les dysphoniques) constitue une source d'information à conserver précieusement. A propos de la réalisation technique, la base de données de parole pathologique SPEEDI-DB est développée dans l'environnement PHP/MySQL sur un serveur Apache avec module sécurisé SSL (https://speedi-db.lpl-aix.fr/physio/). L'anonymisation des données ainsi que l'établissement de consentements éclairés auprès des locuteurs sont des aspects juridiques importants. Concernant l'accès aux données, il est indispensable de gérer des privilèges/rôles accordés aux demandeurs et ce, en fonction des producteurs de données (les hôpitaux). Un modèle de licence d'utilisation a été élaboré de façon à ne pas considérer les cliniciens comme de simples fournisseurs de patients mais au contraire, en les plaçant de façon active dans le processus de recherche. Seule un telle démarche permettra à terme de constituer de vastes corpus de parole "pathologique" multicentriques.

How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers?
Alain Ghio, Gilles Pouchoulin, Bernard Teston, Serge Pinto, Corinne Fredouille et al. Speech Communication, Elsevier : North-Holland, 2012, 54 (5), pp.664-679 <10.1016/j.specom.2011.04.002>

Semi-automatic prosodic annotation for speech corpora

Philippe Martin (LLF, UFRL, Université Paris Diderot)

A dedicated software program integrates prosodic annotation functions which allow the user to graphically represent segments of melodic variations with linear piecewise functions (i.e. one or more straight lines). These linear segments approximate phonetic variations of phonological categories, which are programmable and can be adapted to virtually any phonological model of sentence intonation.
In the incremental prosodic structure model for instance, sentence intonation uses melodic contours categories C0 (terminal falling and low for declarative sentences), C1 (rising above the glissando threshold), C2 (falling above the glissando threshold) and Cn (rising or falling below the glissando threshold). Once the glissando threshold has been defined (Rossi, 1971), the labelling of each segment placed graphically by the user to fit the actual melodic variation, the corresponding phonological category is automatically assigned by the software to the segment, together with the color coding eventually assigned beforehand. Using ToBI notation (Beckman et al., 2005, Delais-Roussarie et al., 2015) such as H*H%, L*L-, , L+H* for F_ToBI is equally easy to predefine and actual acoustic melodic movements are then approximated in terms of targets.
To start the prosodic annotation process, the user has to manually identify the terminal conclusive contour for each sentence of the corpus. From these identifications, the other melodic contours placed on stressed syllables or stressed vowels are automatically assigned to one of the user predefined categories, based on acoustic parameters such as glissando values, duration, etc. All these parameters can be modified by the user.
The user is also helped in the annotation process in difficult recording cases (i.e. low signal to noise ratio, echo, speech overlapping…) by the simultaneous display of the fundamental frequency curve and the fist harmonic of a narrow band spectrogram, using the same frequency scale.
Once the operations of prosodic annotation have been achieved, acoustic parameters of each annotated melodic segment can be transferred to a spreadsheet program in a single mouse click for further analysis. Following this transfer, many statistical tests can be executed from the set of acoustical data, pertaining to the starting and ending points of each linear segment fit on the actual fundamental frequency values: fundamental frequency, intensity and time, as well as segment duration and glissando values. The large set of statistical functions available on spreadsheet programs equipped with suitable statistical packages allows for a detailed investigation pertaining to the validity of the phonological **model implicitly defined in the annotation system selected.**

Quelques formes de l'écriture SMS (eSMS) et traitement semi-automatisé des données

Rachel Panckhurst (Praxiling, Montpellier 3)

Rachel Panckhurst et ses collègues linguistes et informaticiens ont recueilli plus de 90 000 SMS authentiques en langue française auprès du grand public à Montpellier en 2011 (projet *sud4science Languedoc-Roussillon* (http://www.sud4science.org et http://www.msh-m.fr/programmes-2011/sud4science-lr Panckhurst *et al.* 2013), lui-même intégré dans le projet international *sms4science* (http://www.sms4science.org, Fairon et *al*. 2006, Cougnon 2015). Le corpus émanant de cette collecte, *88milSMS*, est téléchargeable sur la grille de services d'Huma-Num (http://88milsms.huma-num.fr/, Panckhurst *et al.* 2014), et sur Ortolang (https://hdl.handle.net/11403/comere/cmr-88milsms, Panckhurst *et al*. 2016). Dans cette communication, elle présentera quelques exemples saillants à partir de la grande diversité de formes existant au sein de *l'écriture SMS* (eSMS) (André 2017, Détrie 2016, Ghliss et Verine 2016, Roche *et al.* 2016), avant d'évoquer le traitement semi-automatisé des données dans une démarche pluridisciplinaire (Panckhurst *et al.* 2016), située entre sciences du langage (Panckhurst et Moïse 2014), informatique et traitement automatique des langues (TAL) (Accorsi *et al*. 2014, Lopez *et al.* 2014, 2015, Panckhurst 2016, Tarrade *et al.* 2017). Elle terminera par l'évocation de pistes ultérieures de recherche (Cougnon *et al*. 2016, Dürscheid et Siever 2017, Panckhurst 2017, Ueberwasser et Stark, 2017).

Bibliographie :

Accorsi P., Patel N., Lopez C., Panckhurst R., Roche M. (2014), « Seek&Hide : Anonymising a French SMS corpus using natural language processing techniques », in *SMS Communication. A Linguistic Approach*, éd L. A. Cougnon, C. Fairon, John Benjamins, Amsterdam/Philadelphia, p. 11-28.

André F. (2017), « Pratiques scripturales et écriture SMS : analyse linguistique d'un corpus de langue française », Doctorat de Sciences du langage, Université Paris-Sorbonne, Jury : C. Fairon, G. Siouffi (directeur), R. Panckhurst, S. Plane, E. Stark. Soutenance le 24 avril 2017. Cougnon L.-A. (2015) *Langage et sms. Une étude internationale des pratiques actuelles*. Presses universitaires de Louvain.

Cougnon L.-A., Maskens L., Roekhaut S., Fairon C. (2016), "Social Media, Spontaneous Writing and Dictation. Spelling variation". In *Journal of French Language Studies.*

Détrie C. (2016).

Détrie C. (2015). Produire du sens en textotant : de quelques innovations lexicales, morphologiques et sémantiques dans les SMS, colloque CINEO 2015, Salamanque, octobre 2015.

Dürscheid C., Siever C. (soumis à *Zeitschrift für Germanistische Linguistik,* 2017), "Jenseits des Alphabets – Kommunikation mit Emojis". Shortened English version, "Beyond the Alphabet – Communication with Emojis", available online: https://www.researchgate.net/publication/315674101_Beyond_the_Alphabet_-_Communication_with_Emojis

Fairon, Cédric, Klein, Jean-René, Paumier, Sébastien (2006), *SMS pour la science. Corpus de 30.000 SMS et logiciel de consultation.* [Manuel+CD-Rom, http://www.smspourlascience.be/], Louvain-la-Neuve : Presses universitaires de Louvain, 2006.

Ghliss, Yosra, Verine, Bertrand (2016) « "Je t'aime fortttttttttt" : la répétition graphémique, marqueur d'émotion dans le genre du discours SMS ? », in A. Krzyzanowska & K. Wolowska (éd.), *Les Émotions et les valeurs dans la communication*, Francfort-sur-le-Main, Peter Lang.

Lopez C., Bestandji R., Roche M., Panckhurst R. (2014) « Towards Electronic SMS Dictionary Construction : An Alignment-based Approach », Actes du colloque *LREC*, Reykjavik, Islande,

26- 31 mai, p. 2833-2838, http://www.lrec-conf.org/proceedings/lrec2014/pdf/753_Paper.pdf

Lopez C., Roche M., Panckhurst R. (2015), « Classification des items inconnus de 88milSMS : aide à l'identification automatique de la créativité scripturale », *Travaux neuchâ telois de linguistique*, Tranel, 2015, 63, 71-86, https://www.unine.ch/files/live/sites/islc/files/Tranel/63/71-86_lopez_al_corr.pdf

Panckhurst R. et Moïse C., (2014), « French text messages. From SMS data collection to preliminary analysis », in SMS Communication. A Linguistic Approach, éd L.-A. Cougnon, C. Fairon, John Benjamins: Amsterdam/Philadelphia, p. 141-168.

Panckhurst R., Roche M., Lopez C., Verine B., Détrie C., Moïse C., (2016b), « De la collecte à l'analyse d'un corpus de SMS authentiques : une démarche pluridisciplinaire », in *HEL*, 38, 2, 63-82, http://www.hel- journal.org

Panckhurst R. (2016), "A digital corpus resource of authentic anonymized French text messages: 88milSMS - What about transcoding and linguistic annotation?" *Digital Scholarship in the Humanities*. Published by Oxford University Press on behalf of EADH. http://dx.doi.org/10.1093/llc/fqw049, 11 pages.

Panckhurst, R. (2017), « Entre linguistique et informatique. Des outils de traitement automatique du langage naturel écrit (TALNE) à l'analyse du discours numérique médié (DNM) », Habilitation à diriger des recherches, Comue Université Paris-Est.

Roche M., Verine B., Lopez C., Panckhurst R. (2016). « La néographie dans un grand corpus de SMS français : 88milSMS ». In : *La neología en las lenguas romá nicas Recursos, estrategias y nuevas orientaciones, Actes du colloque CINEO 2015, 22-24 octobre, Salamanque*. Sous la dir. de Joaquín García Palacios, Goedele De Sterck, Daniel Linder, Nava MarotoMiguel Sánchez Ibáñez et Jesús Torres del Rey. Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation. Frankfurt, Peter Lang.: DOI: http://dx.doi.org/10.3726/978-3-631-69859-4, p. 279-302.

Tarrade L., Lopez C., Panckhurst R., Antoniadis G., (2017) « Typologies pour l'annotation de textes non standards en français », Actes du colloque *TALN*, Orléans, juin, p.118-125, http://taln2017.cnrs.fr/wp- content/uploads/2017/06/actes_TALN_2017-vol2.pdf

Problème de transcription de formes verbales : L'expression du présent (ou du neutre) chez des enfants en Français langue seconde

Jérémi Sauvage et Florence Guiraud (PRAXILING, Montpellier 3)

Les élèves nouvellement arrivés observés en classe sont âgés de 6 à 11 ans. Dans 80% des cas, ils sont déjà plurilingues au moment de leur arrivée en France. Leurs langues d'origine sont l'arabe dialectal, le berbère ou le thaï. Leur parcours migratoire les a amenés dans un autre pays de l'UE auparavant. En d'autres termes, avant d'acquérir le français, ils possèdent donc déjà une première langue seconde : italien, espagnol, hollandais. L'analyse des interactions verbales en situation d'échanges conversationnels, a permis d'étudier un aspect particulier de l'expression des temps verbaux dans ces situations d'interlangue.

Nous posons la question dans cette communication de la difficulté de la transcription et de l'interprétation de formes verbales dans l'analyse d'énoncés comme :

Occurrence N°5 Marco
T : Et ce matin qu'est ce que tu as fait avec MR R… ?
(11)Marco : ce matin //…**on fait** des mathématiques, des mesures, et…/ géographie et voilà//
T : Et du sport aujourd'hui ?
Marco : oui
T : Quel sport ?
(12) Marco : et…./ et…/ Le maitre , y a une batte et une balle , on joue a baseball et ….et …/…ba chercher la balle. (intonation impérative**)**
T : Va chercher la balle ? Qu'est ce que c'est ? Je ne comprends pas ?
(13) Marco:…/: le maitre **tire, …/il fait** comme dé baseball, /…**il fait** avec des battes dures et on va chercher.

Au moment de l'enregistrement, Marco est en France depuis 2 mois. Sa langue maternelle est l'espagnol. Malgré les efforts pour le rassurer, Marco a toujours montré une grande inhibition pour parler français. Ces nombreuses hésitations et moments de silence sont probablement des marques d'un fort sentiment d'insécurité linguistique. Nous nous concentrerons ici sur les extraits de discours suivants: ***On fait*** *des mathématiques (8) Le maitre* ***tire, … /il fait*** *comme dé baseball, /...****il fait*** *avec des battes dures (10)*
Le présent remplace le composé de Passé : sur un fait / il fait. L'indicateur temporel *ce matin* est le seul indicateur du passé. À ce stade de l'apprentissage, l'identification du morphème temporel approprié est difficile pour l'apprenant. L'utilisation du présent apparaît comme une stratégie qui permet à l'orateur de maintenir le thème du processus apporté par cette racine actuelle (Schyler, 1995).

Occurrence n° 6
T : Qu'est-ce que tu veux me raconter ?
(14) *A : Une fois, en Italie,* ***je travaille bien****, et après mon père* ***il m'a donné*** *un moto petit.*

Dans le cas 6, le présent rencontré dans la clause principale est suivi d'un passé (passé composé) dans la deuxième proposition le présent *je travaille bien*, interroge. Nous la transcririons plutôt *je travaill- bien* c'est dire dans une forme verbale complètement neutre, sans flexion temporelle. En effet, ce n'est pas le cas en ce qui concerne le moment de l'énoncé : Nunc (temps absolu), mais c'est comme avant l'autre action (Temps relatif) il m'a donné (Gosselin, 2006). L'indicateur de temps une fois, en Italie, est chargé de spécifier l'interlocuteur (T) ; c'est un indicateur de temps du passé. Du point de vue de son niveau

d'acquisition, l'utilisation du passé composé *m'a donné* atteste que l'apprenant commence à acquérir les formes verbales du passé.

En outre, la temporalité et la modalité sont deux dimensions essentielles de la déclaration. Les études linguistiques, y compris la grammaire, s'opposent généralement à la temporalité et à la modalité des temps. Mais pour Gosselin (2006), la dichotomie entre la dimension temporelle et la dimension modale peut parfois sembler excessive, car elle ne couvre pas toutes les déclarations. En effet, l'utilisation du présent semble être de dimension modale car elle exprime l'attitude d'un orateur vis-à-vis de sa propre déclaration. Il est maintenant verrouillé dans le temps indiquant le passé, évoque une modalité assertive du locuteur. Il semble mettre à jour cette réalité vécue avec une forte charge émotionnelle : *je travaill-bien*, autrement dit, d'un point de vue sémantique : *je travailler bien*.

Nous pouvons encore considérer que l'apprenant et cette forme fixe offrent une plus grande lisibilité sémantique qui permet à l'échange de continuer.

Bibiographie

Gosselin, L. (2006). De distinction entre la dimension temporelle de la modalité et la dimension modale de la temporalité. *Cahiers de Praxématique*, 47, 21-52.

Schlyter, S. (1995). Formes verbales du passé dans des interactions en langue forte et en langue faible. *Acquisition et interaction en langue étrangère*, 6, 129-152.

Segments Répétés pour la veille informationnelle trilingue

Lionel Shen (CLESTHIA, Université Paris 3)

Dans un contexte de sociétés mondialisées, on peut parler de multilinguisme ou encore de plurilinguisme. La traduction devient alors un élément capital pour la communication entre les peuples. Une bonne traduction garantit la qualité de la transmission de toutes les informations. Cependant, devant la gageure que constitue le projet de réaliser une veille multilingue, peut-on utiliser simplement la traduction ?

Notre article tente de mettre en lumière la spécificité et l'efficacité de l'outil Segments Répétés à travers des explorations de corpus thématiques trilingues, français, anglais et chinois, explorations appliquées à la fouille d'informations.

Pour constituer ce travail, deux types de corpus sont mobilisés : un corpus comparable et un corpus parallèle, composés de données textuelles extraites des discours de presse, ainsi que ceux des ONG. La construction de ces deux corpus s'effectue autour de trois thèmes d'actualité ayant pour objet, l'environnement, l'énergie et le changement climatique. Le recueil des deux corpus s'opère à partir des articles de journaux de 1999 à 2014 issus de nos trois sphères de communication, à savoir, le Monde pour la France (4817 articles), le NYT pour les États-Unis (3993 articles) et 1200 médias pour la Chine (14 509 articles).

Après un rappel succinct de l'état de l'art des technicités morphosyntaxiques des trois langues, plus particulièrement sur la segmentation de textes en mots, les corpus sont traités par les méthodes classiques textométriques afin de vérifier la mise en adéquation de ces corpus. Les dépouillements sont réalisés à l'aide des outils de la textométrie, notamment grâce aux analyses factorielles des correspondances (AFC), aux spécificités du modèle hypergéométrique, aux segments répétés ou encore à la carte des sections. Les caractéristiques globales, les convergences et les particularités de ces corpus ont été mises en évidence successivement. Par la suite, le traitement textométrique bilingue a été appliqué sur les mêmes concepts dans l'objectif de mettre en évidence les modes opératoires selon lesquels les corpus multilingues à caractère comparé et parallèle se complètent dans un processus de veille plurilingue.

Dans le cas de corpus comparables, l'outil Segments Répétés permet de dégager des informations clés dans des recherches très spécifiques en français et en anglais, alors qu'en chinois, des valeurs informationnelles non détectables par des moyens traditionnels seront mises à la disposition des utilisateurs.

Obtenus dans la perspective de compléter l'approche additionnelle des explorations multilingues et d'apporter une certaine clarté au discours bilingue de la plaidoirie environnementale des ONG, les résultats de l'étude bilingue (anglais-chinois) des segments répétés parallèles ainsi que leurs analyses montrent que, pour une même information énoncée et décrite en deux langues, la répétition événementielle et thématique est plus saillante en chinois. Pour conclure, l'outil Segments Répétés constitue un atout fondamental pour la fouille d'informations.

Bibliographie :

Bonnafous, Simone, Tournier, Maurice (1995). Analyse de discours, lexicométrie, communication et politique. In : Langages, 29e année, n°117, Paris, Larousse, pp. 67-81.

Habert, Benoît, Nazarenko, Adeline, Salem, André (1997). Les linguistiques de corpus. Paris, Armand Colin/Masson, 254 p.

Habert, Benoît, Zweigenbaum, Pierre (2002). Problèmes épistémologiques : Régler les règles. TAL. Paris, Association pour le traitement automatique des langues, vol. 43, no3, pp. 83-105.

Lafon Pierre (1981). Analyse lexicométrique et recherche des cooccurrences. In: Mots, n°3, octobre 1981. Butor-Rousseau, Péguy, Presse du Zaïre, "la nouvelle droite", vocabulaires, communiste et socialiste, coocurrences? pp. 95-148.

Lafon Pierre, Salem André (1983). L'inventaire des segments répétés d'un texte. In: Mots, n°6, mars 1983. L'oeuvre de Robert-Léon Wagner. Vocabulaire et idéologie. Analyses automatiques. pp. 161- 177.

Lebart Ludovic, Salem André (1995). Statistique textuelle, 342 p.

Salem André (1986). Segments répétés et analyse statistique des données textuelles. In: Histoire & Mesure, 1986 volume 1 - n°2. Varia. pp. 5-28.

Salem André (1987). Pratique des segments répétés. Essai de statistique textuelle, 333 p.

CLARIN ERIC,

L'infrastructure européenne des ressources linguistiques pour les sciences humaines et sociales

Francesca Frontini, Praxiling (Praxiling**,** Montpellier 3), **Christophe Parisse** (MODYCO, Université Paris Ouest Nanterre)

Le premier février 2017 la France a adhèré officiellement CLARIN ERIC, l'infrastructureeuropéenne des ressources linguistiques pour les sciences humaines et sociales.

CLARIN est une fédération de centres de données langagières nationaux, établie pour rendre les ressources linguistiques numériques européennes (écrites, orales ou multimodales) accessibles à travers une authentification unique. CLARIN propose également des outils avancés pour traiter les ressources, où qu'elles soient localisées. Cela est possible grâce à une fédération de centres dans chaque pays membre.

La France a actuellement un statut d'observateur pendant deux ans. Sa participation est coordonnée par la TGIR Huma-Num, avec la participation du consortium CORLI. Ce statut permet aux communautés françaises un accès aux services et aux ressources des pays membres via les identifiants institutionnels et un système de métadonnées commun pour en faciliter le partage et l'interopérabilité.

Une importante retombée pour de la participation à CLARIN est la visibilité internationale pour les ressources françaises déposées dans les centres nationaux; en particulier, les métadonnées d'une ressource sur un centre de pérennisation tel que Ortolang sont maintenant moissonnées par le Virtual Language Observatory (vlo.clarin.eu), le méta-catalogue de CLARIN, et donc plus facilement repérables par des chercheurs étrangers.

Finalement les chercheurs français peuvent participer à d'autres initiatives de dissémination et formation, tels que les séminaires thématiques de CLARIN et les bourses de mobilité, pour des visites de recherche dans des centres CLARIN.

Dans le poster et démo, nous allons donner plus d'information sur CLARIN ERIC; nous montrerons comment accéder à différents corpus écrits et oraux d'autres pays, ainsi qu'à des outils en ligne pour le traitement texte; nous montrerons enfin comment les ressources françaises sont maintenant recherchables à partir du VLO.

The CRFC - Towards a French BNC

Sascha Diwersy (Praxiling Montpellier 3)

The corpus-linguistic analysis of French has been lagging behind other major languages in terms of the diversity and availability of corpora as well as the sophistication of statistical analysis (Deulofeu and Debaisieux 2012: 36). As a result, it has been impossible to produce corpus-based or corpus-driven reference grammars of French or to arrive at a reliable lexico-statistical description of recurrent collocations and colligations in spoken French (Siepmann 2015), to give but two examples.
In our talk, we will discuss the *Corpus de référence du français contemporain* ('reference corpus of present-day French'; henceforth abbreviated as CRFC), a new purpose-built genre-balanced corpus for investigating modern French which has been assembled with a view to remedying the shortcomings of the current situation (see also Siepmann and Bürgel 2014). The CFRC is the first collection of French to put on a par large samples of written text with a substantial amount of spontaneous speech and 'pseudo-spoken' data. The first version of the corpus, which will be extended as new material becomes available, totals 310 million words of the French of France from 1945 to 2014, with more than 90 per cent of texts coming from the last two decades. The corpus is intended to represent the French language as it is relevant to the needs of learners, teachers and researchers in contemporary French language.
Like the British projects *Cobuild* and *BNC* undertaken at the Universities of Birmingham, Oxford and Lancaster in the 1980s and 1990s (cf. Sinclair 1987; Leech 1993), the CRFC promises to be a source for the preparation of corpus-based dictionaries, grammars and language teaching materials. It will accommodate research questions which were impossible to answer using existing corpora, including lexico-grammatical variation across genres and the phraseology of spoken French.

Bibliographie

Deulofeu H.-J., Debaisieux J.-M**.** 2012**.** 'Une tâche à accomplir pour la linguistique française du XXIe siècle : élaborer une grammaire des usages du français.' *Langue française*, 176: 27–46.

Leech G. 1993. '100 million words of English.' *English Today*: 9–15.

Siepmann D. 2015**.** 'Dictionaries and Spoken Language: A Corpus-based Review of French Dictionaries.' *International Journal of Lexicography*, 2: 139–168.

Siepmann D., Bürgel C. 2014. 'Le corpus de référence du français contemporain.' Accessed on 14 November 2015. https://zenodo.org/record/12353?ln=en#.VLebq3ti8Xg .

Sinclair J. 1987. *Looking up: An Account of the COBUILD Project in Lexical Computing*. Birmingham : Collins COBUILD.

La textométrie face à l'analyse de l'oral

Sascha Diwersy (Praxiling, Montpellier 3), **Christelle Dodane** (Praxiling, Montpellier 3) et **Lavie Maturafi** (Praxiling, Montpellier 3)

Dans cette présentation, nous aborderons la question de l'adaptation du format des données orales à un traitement avec des outils destinés au traitement des données écrites. La textométrie (Lebart et Salem, 1994) fournit ainsi des outils qui permettent une classification globale des données textuelles et notamment une classification contrastive. Ce type de traitement pourrait intéresser les différentes communautés de chercheurs travaillant sur l'oral, habitués à une autre méthodologie et d'autres outils, notamment dans le domaine de l'acquisition du langage et de l'analyse conversationnelle. Dans ces domaines, il existe déjà des outils spécifiques et adaptés à leurs besoins. Par exemple, dans l'étude de l'acquisition du langage, le logiciel CLAN (Mc Whinney & Snow, 1990) permet d'effectuer des traitements statistiques tels que la longueur moyenne d'énoncés (Mean Length of Utterance ou M.L.U. Brown, 1973), ainsi que des mesures de diversité lexicale (Type-Token ratio/TTR et VOCD, Malvern & Richards, 1997). Bien qu'elles ne soient pas conçues pour ces domaines, les méthodes textométriques (AFC, Benzécri, 1969 ; calcul de spécificités, Lafon, 1980 ; calcul de co-occurrences, Lafon, 1981 ; temps lexical, Salem, 1988), du fait de leur visée exploratoire et contrastive, permettent de compléter ces méthodes. Nous souhaitons démontrer leur utilité en les appliquant au traitement de corpus oraux, le premier en acquisition du langage (français) et le second, dans le domaine du fonctionnement et de l'appropriation des langues en milieu plurilingue (français-shimaoré). Le premier corpus (Corpus de Paris, Morgenstern & Parisse, 2012[1],[2]) est constitué des productions spontanées d'enfants francophones, enregistrés en situation d'interaction naturelle avec leur entourage et transcrites au format CHAT fourni par le système CHILDES (Mc Whinney & Snow, 1990). Le second corpus (Maturafi, doctorat en cours) est composé des interactions spontanées en shimaoré et en français de locuteurs adultes enregistrées à Mayotte et transcrites en fonction des normes ICOR (Laboratoire ICAR, Université Lyon 2). Les deux corpus ont été transformés pour fournir un modèle de données et un format adaptés au traitement par le logiciel textométrique TXM (Heiden, Magué & Pincemin, 2010). Pour chacun d'entre eux, nous présenterons des exemples d'analyses qui montrent les avantages de l'utilisation de la textométrie dans l'étude des données orales.

Bibliographie

Benzécri, J.P. (1969). Statistical Analysis as a tool to make patterns emerge from data. In Watanabe (eds.). Methodologies of pattern recognition. New-York: Academic Press, 35-74.

Brown, R. (1973). A First language: The early stages. Cambridge: MA, Harvard.

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

Lafon, P. (1980) - Sur la variabilité de la fréquence des formes dans un corpus, Mots N°1, p. 127-165.

Lafon, P. (1981) - Analyse lexicométrique et recherche des cooccurrences, Mots N°3, p. 95- 148.

Lebart, L. & Salem, A. (1994). Statistique textuelle. Paris : Dunod.

MacWhinney, B. & Snow, C.E. (1990). The Child Language Data Exchange System: an update. *Journal of Child Language*, Cambridge, n.17, 457-472.

Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. In A. Ryan and A. Wray (eds), *Evolving models of language. Multilingual Matters*, Clevedon, 58-71.

Morgenstern, A. & Parisse, C. (2012). The Paris Corpus. French Language Studies, 22 (1), 7-12, Cambridge University Press.

Salem, A. (1988). Approches du temps lexical. Statistique textuelle et séries chronologiques, Mots, 17, octobre 1988, 105-143.

[1] http://modyco.inist.fr/data/colaje/ /

[2] http://childes.psy.cmu.edu/access/French/Paris.ht