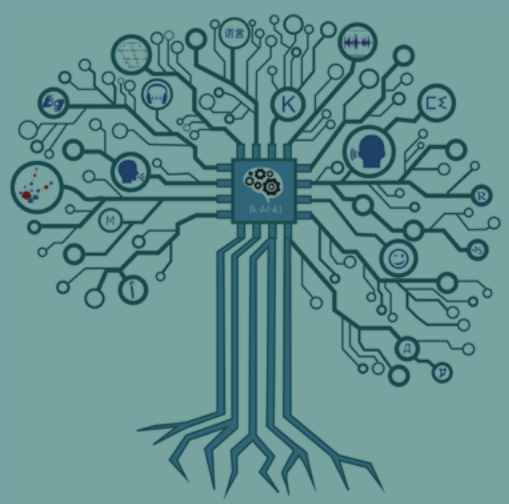
COLDOC 2019 - Fondements empiriques de la linguistique : données de corpus, données expérimentales. Quelles données pour le linguiste ?

Jeanne Conseil, Gasparde Coutanson, Amina Khalfaoui, Yaru Wu, Danilo Lombardo

COLDOC 2019

Colloque des doctorants et jeunes chercheurs du Laboratoire Modyco



Machine Learning & Artificial Intelligence
CC by 2.0 FLICKR

Fondements empiriques de la linguistique : données de corpus, données expérimentales.

Quelles données pour le linguiste ?

27 et 28 novembre 2019

Amphithéâtre W, Bâtiment Max Weber,
Université Paris Nanterre
200 avenue de la république
92001 Nanterre

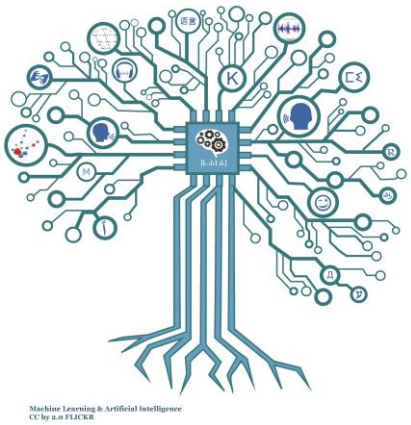
Notre site internet : coldoc2019.frama.site

Notre adresse mail : coldoc2019@gmail.com

Université Paris Nanterre

Université Paris Nanterre
École doctorale Connaissance, langage, modélisation (CLM)

modyco Modèles, Dynamiques, Corpus UMR 7114



COLDOC 2019 - COLLOQUE DE DOCTORANTS ET DE JEUNES CHERCHEURS
MoDyCo UMR 7114 CNRS
Université Paris Nanterre
27 et 28 novembre 2019

Fondements empiriques de la linguistique : données de corpus, données expérimentales. Quelles données pour le linguiste ?

Coldoc est un colloque organisé par les doctorants et jeunes chercheurs en Sciences du langage du laboratoire MoDyCo. Pour sa 13ème édition, Coldoc est centré sur la question des données de la recherche en Science du langage, qu'elles soient collectées dans le cadre d'une linguistique sur corpus ou selon la méthode expérimentale. Nous invitons masterants, doctorants et jeunes chercheurs à venir exposer leurs réflexions sur leurs données de recherche à partir des différentes approches choisies, quel que soit le degré d'avancement de leur recherche. Cet appel à contribution se veut ouvert et cherche à faire dialoguer ces deux approches théoriques afin de s'interroger sur les **apports**, les **limites** et la **complémentarité** possible des **données** résultant de ces deux traditions scientifiques. Cette question traverse différents niveaux d'analyse de la langue (phonétique/phonologique, morphologique, syntaxique, sémantique, pragmatique), différentes unités (lexicale, discursive), différents canaux sensoriels (visuel, auditif, etc) impliquant langue vocale ou signée et différentes modalités (orale, écrite, multimodale).

Dans le cadre de ce colloque, nous souhaitons faire dialoguer la linguistique sur corpus et l'approche expérimentale, ce qui constitue une des dynamiques fondamentales portées par le laboratoire MoDyCo. Des différences méthodologiques entre les deux approches s'observent entre autres au niveau de la formulation d'une hypothèse avant toute collecte et analyse des données, une démarche préférée par les psycholinguistes et qui demeure minoritaire chez les linguistes travaillant à partir de corpus (Gilquin et Gries, 2009), pour des raisons scientifiques mais aussi pour des raisons techniques. Scheer (2004) parle par exemple de *corpus de validation* (avec hypothèse préalable) ou de *corpus heuristique* (sans hypothèse). Pour autant, ces deux approches ne sont pas complètement antagonistes. Plutôt qu'une dichotomie stricte, Gilquin et Gries (2009) proposent un **continuum** entre les deux approches fondé sur l'écologie des données. Par exemple deux jeux de données expérimentales peuvent se positionner différemment sur ce continuum selon qu'elles sont collectées à partir d'activités plus ou moins habituelles pour le locuteur. Surtout, les apports mutuels de ces données permettent de

réactualiser les théories linguistiques. Différentes façons de combiner les deux approches ont pu être envisagées :

1. Le matériel expérimental peut être conçu à partir de données de corpus existantes en les adaptant aux exigences des différents protocoles expérimentaux.
2. Les résultats d'expérimentations peuvent être comparés à des résultats obtenus précédemment sur corpus afin de les valider, nuancer ou infirmer.
3. Les résultats provenant d'analyses de corpus peuvent confirmer, nuancer voire infirmer des résultats provenant d'expérimentations. Par exemple, en phonologie, les données de Liégeois (2014) sur l'acquisition de la liaison valident un modèle théorique fondé sur des données expérimentales.
4. Un linguiste peut encore combiner données de corpus et données expérimentales (ex. De Mönnink, 1997 sur les syntagmes nominaux).

L'ambition de ce colloque est d'apprécier les apports et limites des données collectées par les chercheurs, pour la constitution d'un corpus ou dans un contexte expérimental, dans l'optique de voir comment le choix d'une approche plutôt qu'une autre (analyse de corpus *vs* approche expérimentale) peut influencer les données de la recherche et les résultats. Plusieurs questions, mentionnées ici sans exhaustivité, peuvent être envisagées comme source d'inspiration :

- Quelles sont les points forts et les limites de vos données issues de corpus ou d'expérimentations[1] ?
- Comment les choix méthodologiques de recueil des données influencent-ils vos recherches (données, résultats) ?
- Quels sont les enjeux méthodologiques[2] communs aux deux approches ? En quoi ont-ils un impact sur vos données ?
- Comment l'expérimentateur peut se servir de données et/ou résultats de corpus ? Comment le linguiste travaillant sur corpus peut se servir des résultats d'expérimentations ?
- Comment les outils pédagogiques ou technologiques, développés au terme de votre recherche à partir de vos données, sont-ils influencés par vos choix méthodologiques ?

Notez que les questions présentées ici ne sont proposées qu'à titre d'exemple : vous pourrez n'en traiter qu'une ou présenter tout travail relevant de la méthodologie de corpus, de la méthodologie expérimentale, ou du lien entre les deux.

Au cours de ce colloque, les contributions des participants seront éclairées par des présentations lors de conférences plénières mettant en relation l'approche expérimentale et l'analyse de corpus, en croisant les différents regards. Nous tenterons d'appréhender comment ces deux approches peuvent éclairer mutuellement une même question de recherche. Nous attendons vos propositions de communications orales, démonstrations (20 minutes + 10 minutes de discussion) ou posters sur les thématiques évoquées. Toute autre proposition portant sur la nature et la place des données ou plus largement sur vos recherches en Sciences du langage est encouragée et sera étudiée. Les propositions, d'une page maximum (hors bibliographie) devront

[1] Par *points forts* ou *limites* on entend : écologie, représentativité, contrôle du bruit, variation dans les données, masse des données, coût de la collecte (en terme de temps et de financement), possibilité d'étudier les processus online ou les phénomènes rares, contrôle des variables.

[2] Technologiques, éthiques, juridiques, statistiques, de stockage, réplication des résultats, big data, FAIRisation des données, etc.

être rédigées en français ou en anglais et devront être déposées sur *Easychair* (https://easychair.org/conferences/?conf=coldoc2019) avant le 16 juin 2019.

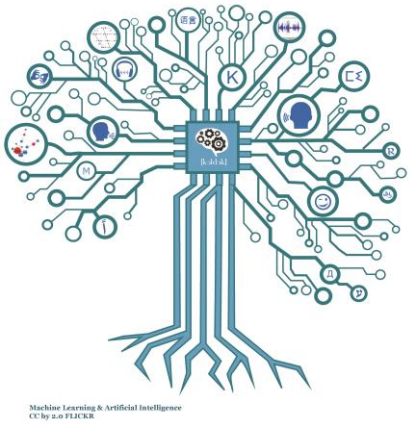
Références :

De Mönnink, Inge. 1997. « Using corpus and experimental data: a multimethod approach ». *Language and computers* 20 : 227-44.

Gilquin, Gaëtanelle, et Stefan Th Gries. 2009. « Corpora and experimental methods: A state-of-the-art review ». *Corpus Linguistics and Linguistic Theory* 5 (1) : 1-26.

Liégeois, Loïc. 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, Clermont Ferrand 2.

Scheer, Tobias. 2004. « Le corpus heuristique : un outil qui montre mais ne démontre pas ». *Corpus*, 3.



COLDOC 2019 - SYMPOSIUM OF PHD STUDENTS AND YOUNG RESEARCHERS

MoDyCo UMR 7114 CNRS

University Paris Nanterre

November 27 & 28, 2019

Empirical Foundations of Linguistics: Corpus Data, Experimental Data. Which data for the linguist?

Coldoc is a symposium organized by PhD students and young researchers in linguistics from the laboratory MoDyCo. For its 13th edition, Coldoc investigates the issue of research data in linguistics, whether collected in a corpus or with the experimental method. Masters students, PhD students and young researchers are invited to reflect on their research data from the different approaches chosen, regardless of the state of advancement of their research. The present call for papers seeks to discuss these two theoretical approaches in order to question the contributions, limits and possible complementarity of the data resulting from these two scientific traditions. This issue concerns several levels of language analysis (phonetic/phonological, morphological, syntactic, semantic, pragmatic), different units (lexical, discursive), different sensory channels (visual, auditory, etc.) involving vocal or signed language and different modalities (oral, written, multimodal).

As part of this symposium, we aim to foster dialog between corpus linguistics and the experimental approach, as this is one of the fundamental dynamics of the laboratory MoDyCo. Methodological differences between the two approaches are observed, for example, at the stage of hypothesis formulation before any data collection and analysis, an approach favored by psycholinguists but less frequent for linguists whose work is corpus-based (Gilquin & Gries, 2009), for both scientific and technical reasons. For instance, Scheer points out that a corpus can have two distinct uses: as a validation corpus (hypothesis-driven) or as a heuristic corpus (without hypothesis). However, these two approaches are not completely opposed. Rather than a strict dichotomy, Gilquin and Gries (2009) propose a continuum between the two approaches based on data ecology. For instance, two sets of experimental data can be positioned differently on this continuum depending on whether they were collected from the speaker's more or less usual activities. Mostly, the mutual contributions of these data make it possible to update linguistic theories. Several ways of combining the two approaches could be considered:

1. The experimental material can be designed from existing corpus data, adapting them to the requirements of the different experimental protocols.
2. The results of experiments can be compared to previously obtained results on a corpus to validate, qualify or invalidate them.
3. The results from corpus analyses can confirm, qualify or even invalidate results from experiments. For example, in phonology, Liégeois' (2014) data on *French liaison* acquisition validated a theoretical framework based on experimental data.
4. A linguist can combine corpus data and experimental data (e.g. De Mönnink, 1997 on *Noun phrases*).

In the present symposium we aim to assess the contributions and limits of the data collected by researchers, either for the constitution of a corpus or in an experimental context, in order to see how the choice of one of the two approaches (corpus analysis vs an experimental approach) can influence research data and the results. Some questions, mentioned below, could be considered as a source of inspiration:

- What are the strengths and limitations of your corpus or experimentation data?[1]
- How do the methodological choices of data collection influence your research (data, results)?
- What are the methodological issues common to both approaches[2]? How do they impact your data?
- How can the experimenter use data and/or corpus results? How can the linguist working on corpora use the results of experiments?
- How are the pedagogical or technological tools -developed at the end of your research based on your data- influenced by your methodological choices?

Note that the questions presented above are only suggested as examples: participants can choose to treat only one or present any work related to the corpus methodology, the experimental methodology, or a combination of the two.

During this symposium, the participants' contributions will be supported by plenaries combining the experimental approach and corpus analysis, in order to tackle different issues from both perspectives. We will try to understand how these two approaches can inform each other about the same research question. We welcome proposals for oral communications, demonstrations (20 minutes + 10 minutes of discussion) or posters on the themes mentioned. Other proposals concerning the nature and place of data or more broadly about your research in linguistics are encouraged and will be studied. Proposals, one page maximum (excluding

[1] By strengths or limitations we mean: ecology, representativeness, noise control, variation in data, mass of data, cost of collection (in terms of time and funding), possibility to study online processes or rare phenomena, control of variables.
[2] Technological, ethical, legal, statistical, storage, results replication, big data, FAIRisation of data, etc.

References), can be in French or English and must be submitted on Easychair (https://easychair.org/conferences/?conf=coldoc2019) before June 16, 2019.

References:

De Mönnink, Inge. 1997. “Using corpus and experimental data: a multimethod approach”. Language and computers 20: 227-44.

Gilquin, Gaëtanelle, & Stefan Th. Gries. 2009. “Corpora and experimental methods: A state-of-the-art review”. Corpus Linguistics and Linguistic Theory 5 (1): 1-26.

Liégeois, Loïc. 2014. “Usage des variables phonologiques dans un corpus d’interactions naturelles parents-enfant : impact du bain linguistique et dispositifs cognitifs d’apprentissage.” PhD Dissertation, Clermont Ferrand 2.

Scheer, Tobias. 2004. “Le corpus heuristique : un outil qui montre mais ne démontre pas”. Corpus, 3.

Conférenciers invités

Gaëtanelle Gilquin (Professeure, Université Catholique de Louvain, Belgique)

Frédéric Isel (Professeur, Université Paris Nanterre)

Anne Lacheret-Dujour (Professeure, Université Paris Nanterre)

Céline Poudat (Maîtresse de conférence, Université Nice Côte d'Azur)

Jacqueline Vaissière (Professeure, Université Sorbonne Nouvelle)

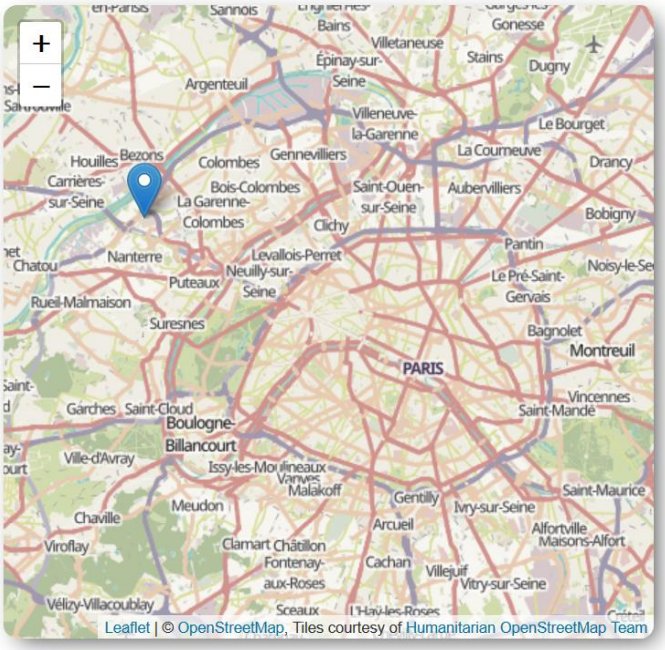
Comité scientifique

Martine Adda-Decker (Université Sorbonne Nouvelle)
Delphine Battisteli (Université Paris Nanterre)
Olivier Baude (Université Paris Nanterre)
Gabriel Bergounioux (Université d'Orléans)
Caroline Bogliotti (Université Paris Nanterre)
Cristelle Cavalla (Université Sorbonne Nouvelle)
Iona Chitoran (Université Paris Diderot)
Chantal Claudel (Université Paris Nanterre)
Sarah de Vogüe (Université Paris Nanterre)
Guillaume Desagulier (Université Paris 8)
José Deulofeu (Aix-Marseille Université)
Iris Eshkol-Taravella (Université Paris Nanterre)
Françoise Gadet (Université Paris Nanterre)
Kim Gerdes (Université Sorbonne Nouvelle)
Béatrice Godart-Wendling (Université Paris Nanterre)
Philippe Gréa (Université Paris Nanterre)
Sylvain Kahane (Université Paris Nanterre)
Takeki Kamiyama (Université Paris 8)
Maria Kihlstedt (Université Paris Nanterre)
Jin-Ok Kim (Université Paris Diderot)
Natalie Kubler (Université Paris Diderot)
Aimée Lahaussois (Université Paris Diderot)
Bernard Laks (Université Paris Nanterre)
Léonardo Lancia (Université Sorbonne Nouvelle)
Sabine Lehmann (Université Paris Nanterre)
Phillippe Martin (Université Paris Diderot)
Aliyah Morgenstern (Université Sorbonne Nouvelle)
Clara Mortamet (Université de Rouen)
Christophe Parisse (Université Paris Nanterre)
Frédérique Sitri (Université Paris Nanterre)
Véronique Traverso (Université Lumière Lyon 2)
Jacqueline Vaissière (Université Sorbonne Nouvelle)

Comité d'organisation

Jeanne CONSEIL (Jeune chercheure postdoctorale)
Gasparde COUTANSON (Doctorante)
Amina KHALFAOUI (Doctorante)
Danilo LOMBARDO (Doctorant)
Yaru WU (Jeune chercheure ATER)

Plan d'accès



L'édition 2019 du ColDoc aura lieu dans le bâtiment Max Weber (W).

Le campus de l'université Paris Nanterre est au 200, Avenue de la République, 92 001 Nanterre.

RER

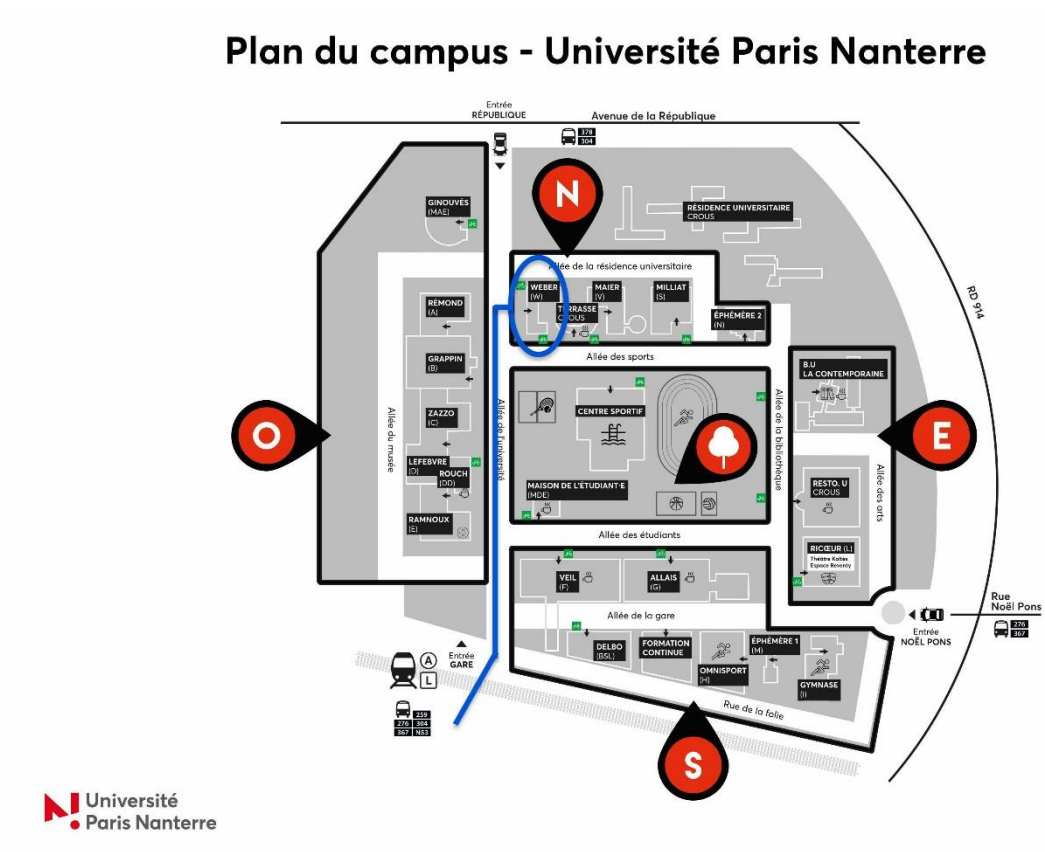
Prendre la ligne A direction SAINT-GERMAIN-EN-LAYE

Descendre à la station NANTERRE-UNIVERSITE

Transilien

Prendre la ligne L depuis GARE SAINT LAZARE, en direction de NANTERRE-UNIVERSITE

Descendre à la station NANTERRE-UNIVERSITE



Résumés des interventions

Présentations orales

Présentations affichées

Un corpus expérimental pour étudier le sourire dans la conversation

Mary Amoyal, Béatrice Priego-Valverde
Laboratoire Parole et Langage, CNRS, Aix-Marseille Université, France.
mary.amoyal@univ-amu.fr, beatrice.priego-valverde@univ-amu.fr

L'objectif de notre étude est de comprendre le rôle du sourire en tant que « facial gesture » [1], ou mouvements conversationnels pendant lesquels l'ajustement des interlocuteurs est nécessaire, dans les transitions thématiques [2] de conversations. Il semble intéressant d'étudier le rôle du sourire dans ces moments particulièrement riches en négociation. Pour ce faire, nous comparons des conversations entre des personnes avec différents degrés de « Personal Common Ground » [3] : dans une condition conversationnelle les interlocuteurs entretiennent des relations amicales, dans l'autre les interlocuteurs ne se connaissent pas. Nous étudions 20 interactions conversationnelles issues de deux corpus : « Cheese ! » [4] recueilli en 2016 réunissant des amis, et « PaCo » recueilli en 2018, une duplication de « Cheese ! » réunissant des inconnus. Les participants sont installés en face à face en chambre anéchoïque. Une caméra est positionnée derrière chaque locuteur, pour les filmer de face, ce qui permet de visualiser en gros plan leurs sourires. Chacun est également appareillé d'un micro-casque pour une bonne capture du signal acoustique tout en préservant la bouche dégagée. Chaque conversation dure en moyenne 15 min et débute par la lecture d'une histoire drôle suivie d'un échange libre. Ce corpus est recueilli dans des conditions expérimentales, tout en permettant aux participants de discuter aussi librement qu'ils le souhaitent. Cette étude a l'ambition d'utiliser une approche de méthodes mixtes qui combine des cadres théoriques qualitatifs issus de l'Analyse Conversationnelle et de la Linguistique Interactionnelle avec des méthodes quantitatives utilisées en Linguistique de Corpus (les schémas d'annotations, les calculs d'accords inter-annotateurs, les modèles mixtes). Dès lors, quels sont les apports et les limites d'un corpus conversationnel semi-contrôlé ? Bien que des biais liés au protocole soient nécessairement présents, ce corpus semi-contrôlé n'empêche en général pas les locuteurs d'oublier leur environnement. Par ailleurs, les enregistrements sur des pistes séparées rendent plus fiables et facilitent la transcription et les annotations. De plus, notre double consigne nous permet d'obtenir pour la tâche de lecture d'une histoire drôle un matériel linguistique semblable pour tous les locuteurs, d'identifier comment se négocie la première transition thématique vers la seconde tâche de conversation et d'analyser la manière dont les locuteurs gèrent leur premier tour de parole [5] vers la séquence conversationnelle, et pour la seconde tâche de conversation sans thème prédéfini d'observer comment les thèmes se négocient selon que les participants se connaissent ou non. A la suite de précédentes analyses du rôle du sourire dans les transitions thématiques [6 ; 7] nous avons mis en place un protocole de détection automatique des sourires avec le logiciel Open Face [8], protocole permettant de vérifier la fiabilité des annotations manuelles et d'élargir les méthodologies existantes [9] concernant l'étude visuelle du sourire. Ce corpus conversationnel semi-contrôlé permet d'approfondir l'analyse des conversations et de ces paramètres multimodaux.

Références

[1] BAVELAS, J. B., & GERWING, J. (2007). Conversational hand gestures and facial displays in face-to-face dialogue. *Social communication*, 283-308.

[2] RIOU, M. (2015). A methodology for the identification of topic transitions in interaction. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*(16).

[3] CLARK, H. H. (1996). *Using Language*. Cambridge University Press.

[4] PRIEGO-VALVERDE, B., BIGI, B., ATTARDO, S., PICKERING, L., & GIRONZETTI, E. (2018). Is smiling during humor so obvious ? A cross-cultural comparison of smiling behavior in humorous sequences in American English and French interactions. *Intercultural Pragmatics, 15*(4), pp. 563-591.

[5] SACKS, H., SCHEGLOFF, E., & JEFFERSON, G. (1974). A Simplest Systematics for the Organization of Turn-taking for conversation. *Language*, 50(4), 696-735.

[6] AMOYAL, M. (2018). *Analyse du sourire lors des transitions thématiques dans la conversation*. Mémoire de recherche de Master 2 en linguistique éxpérimentale , Aix-Marseille Université, Aix-en-Provence.

[7] AMOYAL, M., & PRIEGO-VALVERDE, B. (2019). Smiling for negotiating topic transitions in French conversation. In Angela Grimminger (Ed.). Proceedings of the 6th Gesture and Speech in Interaction (9-15). GESPIN 6, Sept 11-13, 2019, Paderborn : Universitaetsbibliothek Paderborn.

[8] AMOS, B., LUDWICZUK, B., & SATYANARAYANAN, M. (2016). Openface : A general-purpose face recognition library with mobile applications. *CMU School of Computer Science*.

[9] GIRONZETTI, E., ATTARDO, S., & PICKERING, L. (2016). Smiling, gaze, and humor in conversation. Dans L. In Ruiz-Gurillo, *Metapragmatics of Humor : Current research trends* (Vol. 14, p. 235). Amsterdam/ Philadelphia : John Benjamins Publishing Compagny.

Spécificités linguistiques d'un corpus en arabizi

Roya Almostafa
Université Grenoble Alpes - ILCEA4
royaalmostafa@gmail.com

Qu'il soit disponible au marché mondial ou pas encore, arabizi devient une composante majeure de la société de l'information arabe. Ce terme issu de la fusion entre «arab» et «easy» est un style d'écriture moderne, métissé, non conventionnel en chiffres et en caractères latins. Ce langage est surtout utilisé dans les réseaux sociaux et appliqué sur les dialectes arabes. L'absence d'une description précise de ce langage sur le plan linguistique, l'usage libre d'écriture et le contexte non codifié, le rendent peu compréhensible pour certains usagers. Ce problème explique la nécessité d'élaborer des ressources lexicales afin de décrire ce langage sur le plan linguistique. -Objectif de la recherche : Les études scientifiques couvriront l'analyse de différents modes d'expression sur Internet. Ces études se baseront sur nos analyses linguistiques d'un corpus d'expressions en arabizi élaboré et les différents phénomènes linguistiques qui apparaissent. Ceci nous permettra de décrire les caractéristiques linguistiques de ce langage par rapport à l'arabe standard et de proposer des systèmes de conversion entre l'arabe standard et l'arabizi du dialecte libanais dans le but de dessiner sa grammaire. Tracer la grammaire de ce langage et le rendre conventionnelle permettra à une génération d'immigrés arabes, de musulmans et des apprenants de la langue arabe qui ne maitrisent pas l'alphabet de cette langue de réaliser des échanges écrits en se servant de l'arabizi.

A partir de concepts de la linguistique du terrain et du corpus j'ai réussi à construire une base de données de 2000 expressions rédigées en arabizi et appartenant au dialecte libanais en aspirant les données textuelles sur le réseau social Instagram (voir annexe). A l'aide de ce corpus, j'ai réussi à translittérer, traduire et analyser 2000 expressions. J'ai ensuite comparé ces expressions avec l'arabe standard ce qui m'a permis d'observer plusieurs particularités et phénomènes linguistiques tel que le phénomène de contacts des langues que je développerai dans mon intervention.

A partir des statistiques appliquées sur le corpus, les premiers résultats suggèrent une nouvelle formulation de la morphologie du dialecte libanais distinguée de l'arabe standard. -Ex les verbes de la base istaf'3ala / استعفلen arabe standard tel que : istahama, استعفل prennent la base verbale tafa3'ala /عفلtel que tahammam /محت. -Ex le verbe fawwal/ لوف= to full, emprunté à l'anglais est aujourd'hui intégré dans le dialecte libanais avec une forme verbale qui le distingues des autres verbes. -Ainsi, la possibilité d'obtenir de la convention dans une langue considérée non conventionnelle devient fortement possible et envisageable. Observons la lettre ع ayn en arabe qui est transcrite par le chiffre 3 dans presque 75% des mots du corpus.

Extraits du corpus collecté :

Num	Expression arabizi	Translitération	Traduction	Commentaire
1501	Hi kifek ça va ?	Hi كيفك Ça va	Salut, comment vas-tu ? ça va ?	Présence du phénomène de contacts des langues.
10	D3awaTiiik set L kel I Love uu	دعواتك ست الكل I Love you	Je sollicite tes prières ma chère mère, je t'aime.	Set l kel: Expression PO appartenant au dialecte libano syrien signifiant une femme bien respectée. Traduction littérale : Femme supérieur aux autres femmes.
14	allah salmek ya hanoune	الله يسلمك يا حنونة	Que Dieu te garde ma gentille. (Merci tu es gentille)	Allah : formule à deux référents : 1-Sens propre à référent religieux à coder « que Dieu te garde ». 2- Sens figuré lié à l'usage et à la culture du pays
159	Tob el jarra 3a temma btotla3 el benet la emma	طبّ الجرّة عتمّها بتطلع البنت لأمها	Telle mère telle fille.	Expression figées
85	W3i ya mastoule	وعي يا مسطولة	-Ouvre tes yeux, naïve ! -Réveille-toi ! (pour une personne droguée)	Ambiguïté syntaxique

Quand la philologie rencontre le terrain : dialogique rétroactive de corpus et de méthodes

Mathieu Beaudouin
Inalco et Crlao
mathbeaud@gmail.com

La nécessité d'une réflexion méthodologique laisse aujourd'hui rarement la liberté au cher- cheur de s'abstenir d'un positionnement de part ou d'autre de la brèche séparant approches empirico-inductives et hypothético-déductives (Blanchet 2000). Si la nature des données peut parfois enjoindre à pencher plus d'un côté de que de l'autre, des prises de position disciplinaires a priori, opposant généralement, du générativiste au dialectologue, scientificité de la réfutabi- lité (Popper 2002) à complexité des données (Creissels 1979), laissent souvent peu de place au dialogue, au sein d'une discipline partageant un objet pourtant commun, le langage.

Notre travail, la constitution d'une grammaire de référence du tangoute, langue sino-tibétaine médiévale, offre assez étonnamment la possibilité d'emploi d'une méthode complexe [1] autorisant sans effort apparent une coopération harmonieuse des deux approches où, d'une manière assez similaire au mouvement initié par Knorozov pour l'étude du maya, la question de la validation est permise par la multiplication des variétés de corpus. Cet état de fait provient principalement de la dualité de nos données, qui offre l'opportunité d'une dialogique rétroactive féconde (i.e. permettant d'employer les deux méthodes alternati- vement et rétroactivement) entre analyse de textes tangoutes d'une part, et documentation de langues modernes apparentées appartenant au sous-groupe horpa (Sun 2019) du taxon rgyalronguique (Sun 2000) de la famille qianguique de l'autre : au sein de cette dialogique, le caractère heuristique (Scheer 2004) du premier formant de l'analyse fournit des hypothèses à valider à l'autre, et vice-versa. Cet état de fait provient également de la trialité disciplinaire découlant d'un but, la description synchronique d'une langue éteinte, qui ne peut être atteint sans faire ap- pel aux données tant particulières (innovations et rétentions) que nomologiques (panchronie, Hagège & Haudricourt 1978) du comparatisme, ainsi qu'aux implications - respectées ou non - de la typologie (Greenberg 1963). Après une brève présentation de la langue et de l'état de la recherche afférent à sa descrip- tion grammaticale (jusqu'à présent, deux études systématiques : Nishida (1964-66), Kepping (1985), peu de travaux prenant en compte les langues modernes, et toujours aucune grammaire de référence), on se recentrera sur la présentation de nos corpus et du processus de constitution de ces derniers, avant d'essayer d'étudier en détail le fonctionnement de cette boucle rétroac- tive, en illustrant notre propos à l'aide d'exemples concrets.

1. Dans le sens où elle permet une pensée complexe (Morin 1986), i.e. une stratégie de raisonnement où l'empirisme se nourrit de l'hypothèse, qui se nourrit elle même de l'expérience.

Références

[1] BLANCHET, P. *La linguistique de terrain. Méthode et théorie. Une approche ethno- sociolinguistique*. Presses universitaires de Rennes. (2000).

[2] CREISSELS, D. *Unités et catégories grammaticales : réflexions sur les fondements d'une théorie générale des descriptions grammaticales*. Publications de l'Université des langues et lettres de Grenoble. (1979).

[3] GREENBERG, J. Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In Universals of Language (1963), 73–113. London : MIT Press.

[4] HAGÈGE, C. AND HAUDRICOURT, A.-G. *La phonologie panchronique*. PUF. (1978).

[5] KEPPING, K. Тангутский язык : Морфология. [The Morphology of the Tangut Language]. Москва : Наука [Moscow : Nauka]. (1985).

[6] MORIN, E. *La méthode. Livre 3 : La connaissance de la connaissance*. Seuil. (1986).

[7] NISHIDA, T. 西田龍雄. 西夏語の研究 — 西夏語の再構成と西夏文字の解読 *Seikago no kenkyū - Seikago no saikōsei to seika moji no kaidoku [Tangut Studies - Decipherment of the Tangut Script and Reconstruction of the Tangut Language]*. 座右宝刊行会 Zauho Kankokai [2 volumes]. (1964-66).

[8] POPPER, K. *The Logic of Scientific Discovery*. Psychology Press. (2002).

[9] SCHEER, T. Le corpus heuristique : un outil qui montre mais ne démontre pas. *Corpus, 3*. (2004).

[10] SUN, J. T.-S. Parallelisms in the Verb Morphology of Sidaba rGyalrong and Lavrung in rGyalrongic. *Language and Linguistics* 1, 1 (2000), 161–190.

[11] SUN, J. T.-S. *The Ancestry of Horpa : Further Morphological Evidence*. (2019) (forthcoming).

Transcription formelle et capture de mouvement : vers une modélisation des constituants gestuels de la forme des signes

Léa Chevrefils
Université de Rouen
lea.chevrefils@univ-rouen.fr

Les Langues des Signes (LS) répertoriées aujourd'hui sont au nombre de 142 [1] et concernent plusieurs dizaines de millions de locuteurs dans le monde. À la fin du XXe siècle William Stokoe [2] met en exergue la double articulation des LS permettant de les intégrer dans le champ de la linguistique. Malgré cela, une forte dissymétrie persiste au sein des études scientifiques des LS face à celles des langues vocales (LV). Ce manque de ressources est en partie dû à une absence de système de transcription standardisé : l'annotation de corpus se fait principalement par « ID-gloses » [3], procédé s'apparentant plus à de l'identification par traduction qu'à de la transcription proprement dite. Pour parvenir à une modélisation des LS stable, il est fondamental d'accéder au préalable à une segmentation automatique des signes, d'une part, et, d'autre part, à recourir à une transcription standardisée. Or, la segmentation se confronte à la fluence du mouvement, omniprésente au sein d'une locution en LS. Mon projet de thèse propose d'aborder cette problématique en précisant les rapports de couplage moteur qu'entretiennent les deux paramètres de l'emplacement et du mouvement. Les retombées attendues relèvent, d'un point de vue théorique, d'une modélisation formelle de la langue et, d'un point de vue plus appliqué, de la mise en place d'algorithmes de suivi des mouvements des signes. Afin de traiter efficacement ces questions de structuration profonde des LS, un corpus de locuteurs en LSF est enregistré par une caméra vidéo ainsi que par un système de captation 3D modulaire offrant la possibilité d'étudier la cinématique du mouvement. L'exploitation de ce corpus se fait en deux phases successives : tout d'abord la vidéo est annotée avec la police de caractères Typannot, correspondant à une transcription orthographique ; ensuite, les résultats sont comparés avec les données physiologiques du corpus en MoCap — données de type phonétique —, afin d'établir formellement les liens qu'entretiennent les deux paramètres étudiés. Cette opportunité, celle de nous tourner vers ces nouvelles technologies en linguistique des LS, soulève de nouvelles questions sur la nature des informations qui nous parviennent. Il est impossible d'accéder aux données brutes d'un enregistrement en MoCap avec autant de facilité qu'avec celles enregistrées en vidéo. Le Neuron enregistre 60 fois par seconde des valeurs sous la forme de positions relatives, et ce, selon les trois axes X Y et Z de chaque segment du corps. Les informations fournies s'apparentent à une longue suite de nombres, pour laquelle on ne peut faire abstraction d'un temps de traitement qui précède leur exploitation. Or, si la définition des données primaires tient à leur transparence, à l'immédiateté de leur accès, nous sommes peut-être confrontés à une redéfinition des frontières distinguant les données brutes des données secondaires.

Références

[1] Ethnologue, Languages of the World [en ligne]. Disponible sur : https ://www.ethnologue.com/subgroups/sign-language

[2] STOKOE, W. C. Sign Language Structure : An Outline of the Visual Communication Systems of the American Deaf. *Journal of Deaf Studies and Deaf Education* 10, 1 (2005), 337.

[3] JOHNSTON, T. From archive to corpus : Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics* 15, 1 (2010), 106–131.

L'élaboration d'un corpus de « petites phrases » politiques : particularités et richesse du questionnement

Damien DEIAS
Crem, Université de Lorraine, F-57045
deiasdamien@yahoo.fr

La diffusion médiatique des petites phrases est devenue un phénomène central de la communication politique, suscitant d'abondants commentaires des journalistes. Bien que les petites phrases ne soient pas apparues au XXIe siècle, certaines étant demeurées célèbres au cours des siècles, ainsi que « *l'État, c'est moi* » qu'aurait prononcé Louis XIV en 1655, leur multiplication récente a suivi le développement de nouveaux moyens de communication et de la nécessaire adaptation du monde politique à cette évolution. A chaque événement et crise de la démocratie ses petites phrases. Mon travail de doctorat, sous la direction du Professeur Mustapha Krazem, se propose de mener une étude générale de ces « phrases sans texte » dans le domaine de la politique, selon la formule de Dominique Maingueneau [1]. Comment naît une petite phrase ? Quelles sont les conditions de son détachement ? Existe-t-il des récurrences morpho-syntaxiques ? Comment agissent-elles sur le monde ? Sont-elles l'amorce d'une nouvelle rhétorique ?

La particularité de l'objet étudié au regard d'autres formes de discours réside dans ses conditions de production, qui ne peuvent être envisagées du seul point de vue du locuteur. Les phénomènes de reprises, de partages, de modifications de la formulation, du sens font pleinement partie de la définition de ce genre, instaurant une scène d'énonciation. Dès lors, pour le chercheur, la constitution d'un corpus qui ne peut être seulement linéaire, mais qui collecte des éléments pertinents en aval et en amont de la production des petites phrases. Le corpus se constitue de petites phrases collectées dans la presse écrite et audiovisuelle et sur les réseaux sociaux, en particulier Facebook et Twitter, mais également de commentaires de journalistes sur celles-ci afin d'en apprécier la réception médiatique. Outre le contexte, le cotexte est une donnée déterminante puisque les petites phrases sont des énoncés détachés. Les conditions de détachement (aphorisation, position dans le discours ...) déterminent donc le choix du segment susceptible d'être retenu. En aval, la réception [2], les divers phénomènes de reprises, de contextualisation, de transformations en mèmes, nécessitent également une collecte rigoureuse et adaptée. La constitution opératoire d'un tel corpus constitue donc un défi pour le chercheur, en même temps qu'elle permet de mettre en valeur la richesse de l'objet et des questionnements qu'il soulève : est-ce l'acte de détachement d'une petite phase qui définit ce type de discours ? Tout type d'énoncé peut-il devenir devenir une petite phrase ?

Références

[1] MAINGUENEAU D. *Les Phrases sans texte* (2012), 12-70.

[2] LE SÉAC'H, M. *La petite phrase* (2015), 15-21.

Analyse sur corpus comparable de la métaphore conceptuelle dans le discours spécialisé

Charlélie Fanget
Université Paris Diderot, Paris 7
fanget.c@gmail.com

Depuis plus de quarante ans, la théorie de la métaphore conceptuelle (Lakoff et Johnson, 1980) s'est imposée comme l'approche dominante pour aborder les phénomènes métaphoriques. Parallèlement, elle n'a jamais été aussi nuancée voire critiquée que depuis l'émergence et le développement de la linguistique de corpus.

En effet, il a souvent été reproché à la théorie de la métaphore conceptuelle de ne s'appuyer que sur l'intuition du locuteur natif ainsi que sur des exemples ad hoc pour formuler ou valider des hypothèses (Stefanowitsch et Gries, 2008). Cela aurait pour conséquence de mettre de côté certains aspects irréguliers des expressions métaphoriques, et de ne travailler que sur une collection appauvrie de métaphores (Kövecses, 2011). Dès lors, la collecte et l'analyse de données attestées de la langue, recueillies dans des contextes naturels, permettrait aux chercheurs d'atteindre une validité écologique (Low et al, 2010).

Ainsi, cette communication a pour ambition de présenter une méthodologie de détection semi- automatique des phénomènes métaphoriques au sein de deux corpus comparables français et anglais. L'accent sera mis d'une part sur l'aspect heuristique de la méthode, dans le sens où aucune théorie ni hypothèse ne préside aux observations sur corpus, et d'autre part sur l'authenticité des données recueillies.

Puisque la métaphore conceptuelle n'est jamais associée à une forme linguistique précise, il serait nécessaire de posséder des connaissances préalables sur les phénomènes métaphoriques au sein d'un corpus, ce qui exclurait a priori toute forme d'analyse sur corpus heuristique (Scheer, 2004). Or, en se basant sur une analyse de candidats-termes et de marqueurs discursifs spécifiques, il est possible de mettre en évidence de grandes tendances métaphoriques ; cela ouvre ainsi la porte à des approches "corpus-driven" de la métaphore (Tognini-Bonelli, 2001).

Dans un premier temps, nous présenterons les contraintes et les impératifs auxquels le corpus devait répondre. Ensuite, nous dévoilerons la méthodologie employée pour détecter des éléments métaphoriques dans le corpus, en insistant sur les forces et les faiblesses d'une telle approche. Enfin, nous présenterons des exemples de métaphores conceptuelles provenant du corpus.

Notre communication se basera sur le corpus comparable que nous avons constitué lors de nos travaux de thèse de doctorat, qui porte sur l'influence du genre discursif sur la métaphore conceptuelle dans le discours financier.

Références

[1] KÖVECSES, Z. Methodological issues in conceptual metaphor theory. *Windows to the mind : Metaphor, metonymy and conceptual blending* (2011), 23-40.

[2] LAKOFF, G. AND JOHNSON, J. *Metaphors We Live By, Chicago* : University of Chicago. (1980).

[3] LOW, G., DEIGNAN, A., CAMERON, L., TODD, Z. *Researching and applying metaphor in the real world* (Vol. 26). John Benjamins Publishing. (2010).

[4] SCHEER, T. Le corpus heuristique : un outil qui montre mais ne démontre pas. *Corpus*, 3 (2004).

[5] STEFANOWITSCH, A. AND GRIES, S. T. (Eds.). *Corpus-based approaches to metaphor and metonymy*, 171 (2008). Walter de Gruyter.

[6] TOGNINI-BONELLI, E. *Corpus Linguistics at Work*, (2001). Amsterdam : John Benjamins.

Conducting experimental research on oral languages : the case of the study of information structure in Gizey

Guillaume Guitang
Université libre de Bruxelles
Guillaume.Guitang@ulb.ac.be

This presentation will address methodological issues in the study of information structure (focus, topic, voice, given etc.) in an unwritten and under-described language with little or no available and usable corpora. It is part of a broader project to describe the grammar of Gizey, a language spoken by some 10 to 12 000 speakers (Ajello 2006, De Dominicis, 2008) mainly located in northern Cameroon. The study is based on experimental tasks laid out in the Questionnaire on Information Structure (hereinafter, QUIS). QUIS experimental tasks, aim at eliciting "spontaneous sentences or short dialogues with specific information structural content" (Skopeteas, Fiedler, Hellmuth, Schwarz, Stoel, Fanselow, Féry & Krifka, 2006, p. 6) ; with the use of pictures, short films, games etc. Language consultants have to describe situations, narrate events, perform short games and answer specific questions based on non-verbal stimuli. These tasks are organised into four field sessions where various conditions (all new, given patient, inanimate agent etc.) are tested with a primary and secondary consultant. Data resulting from these experiments have to be annotated and distributed online to serve potentially as corpora for future research. My contribution will focus on describing the recent field experience with collecting experimental data with the QUIS (roughly 7 hours of unedited recordings) and discussing the main challenges (recruiting suitable consultants, preparing and manipulating stimuli and instructions etc.) and limitations (degree of naturalness of the data, appropriateness of stimuli, etc.) relating to this methodological choice.

Références

[1] AJELLO, R. The importance of having a description of the endangered languages. In A. De Dominicis (Ed.), *Undescribed and Endangered Languages the Preservation of Linguistic Diversity* (2006), 8-20. Newcastle : Cambridge Scholars Press.

[2] DE DOMINICIS, A. Phonological Sketch of Gizey. *Studi Linguistici e Filologici Online* 6 (2008), 1-78.

[3] SKOPETEAS, S., FIEDLER, I., HELLMUTH, S., SCHWARZ, A., STOEL, R., FANSELOW, G.,... KRIFKA, M. Questionnaire on Information Structure (QUIS) : Reference manual. *Working Papers of the SFB 632* 4, (2006).

Étudier des phrasèmes du berbère : pourquoi choisir l'entretien semi- directif et l'observation participante ?

Rabih Driss Lacnad
Inalco, USPC, Paris
drissrabihprof@gmail.com

Les discussions relatives aux approches des enquêtes se focalisent sur les frontières entre les conceptions subjectives des enquêteurs lors de l'interprétation des phénomènes et l'obligation de la recherche méthodologique de la connaissance scientifique. Si les approches, dites statistiques ou quantitatives, sont utilisées par les chercheurs en quête d'« objectivité », l'approche qualitative ou enquête de type ethnographique se rapporte à la notion d'authenticité, et « se veut au plus près des situations naturelles des sujets – vie quotidienne, conversations -» (Olivier de Sardan, 1995). Ces techniques permettent, comme le décrit Olivier de Sardan (1995), de « produire des connaissances in situ, contextualités, transversales, visant à rendre compte du point de vue de l'acteur, des représentations ordinaires, des pratiques usuelles et (…) leurs significations autochtones » (Olivier de Sardan, 1995). Quant à notre recherche de thèse consistant en l'étude thématique et sémantico-syntaxique des expressions idiomatiques berbères – cas du parler des Ayt Ḥmad Uɛisa, Maroc central-, et dont l'analyse porte sur un corpus oral constitué de séquences figées, nous avons opté pour deux techniques d'enquête : l'entretien semi-directif et l'observation participante. Dans notre communication, nous allons justifier ce choix méthodologique en expliquant comment ces deux techniques peuvent être utilisées dans le domaine de la phraséologie. Il importe de noter que, lors de la collecte des expressions idiomatiques, nous avons pris en considération la nature de l'objet de thèse, c'est-à-dire, comment l'usage de ces deux techniques peut être au service de l'analyse sémantique, thématique et syntaxique que nous menons dans notre travail : contextualiser l'usage des expressions, différentes formes syntaxiques que prend chaque expression.... Nous démonterons, à partir des exemples concrets, comment l'usage l'entretien semi-directif et l'observation participante dans le recueil des données influencent notre recherche. Nous aborderons également les points forts et les limites de nos données issues de ce corpus.

Références

[1] OLIVIER DE SARDAN, J.-P. « La politique du terrain. Sur la production des données en anthropologie ». *Enquête. Archives de la revue Enquête*, 1 (1995), 71-109.

La correction des enseignants en contexte scolaire : analyses d'un corpus multimodal écrit, audio et vidéo

Arnaud Moysan
Laboratoire Clesthia (EA 7345)
arnaud.moysan@sorbonne-nouvelle.fr

L'objectif de cette communication est de proposer une réflexion sur la multimodalité de notre corpus de recherche qui, partant d'un geste professionnel présent sur des copies d'élèves, nous a conduit à considérer le contexte qui entoure l'activité de correction, embrassant de fait les différents acteurs de celle-ci, et nous obligeant à ouvrir notre recueil de corpus aux données vidéo (enregistrements de classes) et audio (entretiens semi-directifs menés auprès de collégiens).

La correction des écrits scolaires est une activité chronophage à laquelle tout enseignant se confronte et ce quel que soit son niveau d'enseignement, de la classe préparatoire à l'université. Pour autant la recherche ne s'est que très rarement intéressée au sujet. En 2008, Jean-Luc Pilorgé soutient sa thèse sur la correction des enseignants, laquelle repose en partie sur l'analyse linguistique des corrections effectuées par des enseignants stagiaires de français. Néanmoins, le corpus de Pilorgé demeure assez éloigné des pratiques réelles de correction au collège, puisqu'il repose exclusivement sur des écrits de néo-enseignants, oblitérant de fait le vaste panel des enseignants plus expérimentés.

C'est pourquoi notre présente recherche s'est donnée pour objectif le recueil et l'analyse d'un corpus d'écrits scolaires actuel, écologique, constitué auprès de classes de 6^{e} et de 3^{e}. Notre corpus actuel s'élève à environ 1000 copies d'élèves – de type rédactions de français – qui sont, par la suite, transcrits numériquement afin de permettre quantitatives. Les séances durant lesquelles les élèves ont écrit ces rédactions ont également été filmées de manière à recueillir les interactions orales entre l'enseignant et les élèves, lorsque ces derniers interrogent le premier sur l'activité d'écriture.

Cependant, l'analyse seule des copies d'élèves n'a pas été jugée suffisante en ce qu'elle éclipse un élément des plus importants dans la correction : la prise en compte du destinataire de celle-ci, c'est-à-dire l'élève. En effet, il nous est apparu évident de donner voix à celui-ci et d'analyser son discours sur la correction, jugée par les enseignants suivis cette année comme étant un outil d'aide à l'écriture. Mais ce discours de l'enseignant sur la correction est-il le même chez l'élève ? N'est-il pas au contraire sujet à « malentendus » (Bautier, Rayou, 2012) entre les deux ? Ces malentendus constituent-ils un frein dans l'apprentissage de l'écriture textuelle chez les élèves ? Afin d'obtenir ce précieux point de vue, nous avons mené des entretiens audio auprès d'élèves dans le but d'obtenir leur représentation sur l'écriture (en général) et la compréhension des formes constitutives de la correction de leur enseignant (en particulier).

Notre corpus multimodal est donc vaste, ce qui pose certains problèmes dans l'analyse de celui-ci. Il demeure néanmoins important car il permet de rendre compte de situations réelles de classe en ce qui concerne l'écriture scolaire, vecteur important des inégalités dans les apprentissages.

Références

[1] BAUTIER, E. AND RAYOU, P. *Les inégalités d'apprentissage : Programmes, pratiques et malentendus scolaires*, PUF, Paris. (2009).

[2] BUCHETON, D. « Les postures de lecture des élèves au collège ». *Enseigner la littérature* (2000), 201-14.

[3] PILORGE, J.-L. *Un lieu de tension entre posture de lecteur et posture de correcteur : les traces des enseignants de français sur les copies des élèves. Traces de lectures/traces de corrections sur des écrits fictionnels de collégiens*. (2008).

Les régulateurs discursifs en français et en japonais : une étude contrastive.

Chiara Manno
Université Paris Nanterre
cmanno@parisnanterre.fr

Notre recherche s'inscrit dans le domaine de l'analyse des interactions verbales. Par « interaction verbale », on entend « toute forme de discours produit collectivement par l'action ordonnée et coordonnée de plusieurs interactants » [1]. L'entrée d'analyse est le comportement de l'allocutaire, et plus précisément les signaux verbaux et non verbaux qu'il emploie pour montrer sa participation active dans l'interaction, autrement appelés régulateurs discursifs. En nous appuyant sur les études antérieures [2, 3, 4], la définition de « régulateur discursif » que nous adoptons est la suivante : il s'agit de signaux d'écoute employés par l'allocutaire pendant le tour de parole du locuteur pour satisfaire des fonctions communicatives de contact, de perception, de compréhension, d'approbation et d'attitude. On part de l'hypothèse que la fréquence, les fonctions et les variétés des régulateurs discursifs sont spécifiques à la culture des interactants [5] ; les cultures et les langues étudiées dans notre thèse sont le français et le japonais. Des recherches ont montré que les régulateurs discursifs japonais s'emploient plus fréquemment que les régulateurs discursifs de langues comme l'anglais par exemple [6, 7, 8] ; toutefois, il n'existe aucun travail comparatif entre le japonais et le français. Nous comblerons cette lacune à travers une étude contrastive basée sur l'analyse de deux corpus issus de ces deux langues. Le type d'interaction étudiée est la conversation entre amis, souvent caractérisée par des récits et des séquences explicatives où l'un des interactants raconte un événement et l'autre joue le rôle de récepteur du discours [9]. Nous avons personnellement enregistré le corpus japonais pendant une année de recherche effectuée à l'Université de Kobe (Japon) : douze étudiants ayant entre 18 et 20 ans ont conversé librement entre eux en dyades pendant vingt-cinq minutes environ. Les conversations ont été filmées dans le but de tenir également compte des régulateurs non verbaux tels que le regard et les mouvements de tête ; nous sommes en train de commencer le recueil du corpus français auprès de l'Université Paris Nanterre en utilisant la même méthodologie. L'analyse quantitative des données, effectuée à travers le calcul statistique Lmer (*Linear mixed- effects models*), nous permettra de comprendre s'il y a des différences significatives dans la fréquence des régulateurs discursifs employés par les locuteurs français et celle des régulateurs discursifs employés par les locuteurs japonais. L'analyse qualitative, basée notamment sur l'approche méthodologique de l'Analyse conversationnelle, nous permettra de saisir les variétés morphosyntaxiques des régulateurs français et japonais, leur place dans l'organisation séquentielle de l'interaction et leurs fonctions. Avec cette recherche, on veut fournir des éléments de réflexion sur les langues japonaise et française en vue d'une meilleure compréhension des deux communautés. Les résultats pourront en outre constituer des pistes pour l'élaboration de supports didactiques pour l'enseignement des stratégies conversationnelles orales.

Références

[1] KERBRAT-ORECCHIONI, C. « La notion d'interaction en linguistique : origine, apports, bilan », *Langue Française* 117 (1998), 51-67.

[2] ALLWOOD, J., NIVRE, J., AHLSEN, E. . « On the semantics and pragmatics of the linguistic feedbacks », *Semantics* 9, 1 (1993), 1-26.

[3] KERBRAT-ORECCHIONI, C. *Les interactions verbales*, Tome I, Paris, Armand Colin. (1990).

[4] YNGVE, V. « On getting a word in edgewise », *Papers from the Sixth regional Meeting of the Chicago Linguistic Society* (1970), 567–577.

[5] HEINZ, B. « Backchannel responses as strategic responses in bilingual speakers' conversations », *Journal of Pragmatics* 39 (2003), 1113-42.

[6] CLANCY, P., THOMPSON, S., SUZUKI, R., TAO, H. (1996). « The conversational use of reactive tokens in English, Japanese, and Mandarin », *Journal of Pragmatics* 26 (1996), 355-387.

[7] MAYNARD, S. « On back-channel behavior in English and Japanese conversations », *Linguistics*, 24, 6 (1987), 1079-1108.

[8] ŌHAMA, R. *Nihongo kaiwa ni okeru tān kōtai to aizuchi ni kan suru kenkyū* (« Recherches sur les tours de parole et les régulateurs discursifs dans la conversation japonaise »), Hiroshima, Keisuisha. (2006).

[9] TRAVERSO, V. L'analyse des conversations. Paris, Armand Collin. (2013).

Gathering insights into peripheral features of grammar.

Cameron Morin
Ecole Normale Supérieure Paris-Saclay
cameron.morin@neuf.fr

This paper focuses on some empirical problems and solutions in the study of rare language variation, through the case study of dialect syntax in English, and drawing on substantial fieldwork by the author.

Multiple modals are peripheral but noticeable constructions in several British and American basilects. The examples (Annex 1, by no means exhaustive) come from Borders Scots (Miller 2004).

Investigating these features is an empirical and methodological challenge. Firstly, classic corpora-based enquiries reveal themselves to be insufficient : for instance, the large amount of data available in the corpora of the Angus McIntosh Centre for Historical Linguistics (AMC, University of Edinburgh) yields next to 0 frequency hits for multiple modals. This is supposedly due to the marginality of the structures, even in the varieties where they have been suggested to occur. One possible solution is the compilation of corpora from new sources where oral speech and dialects seem to thrive in written form – one of these is Twitter (see, for instance, the Tweetolectology project at Cambridge).

Alternative sources of data collection may prove more useful, such as fieldwork experimentation directly interacting with the speech communities concerned. In January 2018, the author conducted a field experiment in the town of Hawick (Borders), distributing a questionnaire to approximately 60 respondents from various age groups and occupations. The questionnaire was semi-structured, and revolved around tasks of judgment elicitation and syntactic manipulations to get a quantitative and qualitative picture of double modals in this representative locus of Borders Scots (these will be described in more detail during the presentation) which could never have been provided through a classic corpus.

However, do intuition-based judgments unfailingly deserve our full trust (Labov 1996) ? There are serious empirical issues with these alternative methods, and close scrutiny must be brought to the ways of avoiding their biggest pitfalls, notwithstanding their need to contend with the infamous “observer’s paradox”.

These new problems might be well compensated, however, by a new combinatorial and multidimensional approach to rare dialect syntax, by reappraising both corpora compilation and fieldwork ; and cross-examining specific quantitative and qualitative aspects of their individual components to establish the coherence of the resulting picture. In order to strengthen the results of questionnaires, fine-grained acceptability judgment methods updated with respect to their psycholinguistic validity will also be considered. Method triangulation is a view on the rise in studies of language variation and change (Krug & Schlüter 2013), and it is one the author is currently developing for his doctoral investigation of multiple modals in English and their syntactic, semantic, and pragmatic implications.

Annex 1

*(a) He'**ll can** help us the morn.*
*(b) They **might could** be working in the shop.*
*(c) She **might can** get away early*
*(d) Wi his sair fit he **would** never **could** climb the stairs.*

Références

[1] Angus McIntosh Centre for Historical Linguistics. Scots and Scottish English corpora. Edinburgh : University of Edinburgh. http ://www.amc.lel.ed.ac.uk/ ?page_id=1769 (accessed 12/06/19).

[2] MILLER, J. The morphology and syntax of Scottish English. In Schneider, Edgar & Berns Kortmann (eds.), *A Handbook of Varieties of English*, Berlin : Mouton de Gruyter. (2004).

[3] LABOV, W. When intuitions fail. In McNair et al. (eds.), *Papers from the Parasession on Theory and Data in Linguistics* 32 (1996), 77–106.

[4] KRUG, M. & SCHLÜTER, J. (eds.). *Research methods in language variation and change*. (2013). New York : Cambridge University Press.

Combiner l'analyse interactionnelle et l'approche expérimentale dans la recherche sur la compréhension de l'interaction en français L2

Simone Morehed
Université de Fribourg
simone.morehed@unifr.ch

L'importance des compétences interactionnelle et pragmatique, outre celles linguistiques, est de plus en plus prise en compte dans la recherche et dans l'enseignement en L2, l'interaction étant l'outil par lequel nous accomplissons la plupart des activités de la vie quotidienne et donc d'une importance primordiale dans l'acquisition d'une L2 (Gülich & Mondada 2001).

Les études existantes dans ce domaine sont surtout de deux catégories. Premièrement, il y a celles visant la production en interaction en L2, avec des données majoritairement authentiques, c'est-à-dire des interactions produites de manière spontanée par les locuteurs, sans des instructions précises. La compréhension n'est incluse qu'en tant qu'affichée par la production (Oursel 2013).

Deuxièmement, il existe des études visant la compréhension pragmatique par des méthodes expérimentales, mais elles font surtout usage d'un matériel non-authentique, c'est-à-dire inventé dans un but scientifique ou didactique. Ce type de matériel ne contient pas les caractéristiques propres à l'interaction, p.ex. des hésitations et des répétitions (Bloomfield et al. 2010). De plus, l'input ainsi que la tâche à effectuer par les participants sont souvent en format écrit, bien que les situations décrites soient censées d'illustrer une interaction orale authentique (Taguchi 2015).

En fait, nous savons que les apprenants L2, même au niveau avancé, s'expriment de manière différente des locuteurs L1, mais nous ne savons pas ce qu'il y a en interaction qui leur pose problème (Bardovi-Harlig & Salsbury 2004).

Dans cette présentation, nous discuterons les défis et les possibilités d'utiliser l'analyse conversationnelle et une approche expérimentale, dans le but de viser la compréhension de l'interaction (Kendrick 2017).

Nous présenterons une étude pilote récemment effectuée, où nous avons visé la compréhension de l'interaction chez des apprenants avancés en français L2. Des extraits audio d'interactions authentiques sont présentés aux participants, et ils répondent ensuite à des questions sur des échelles Likert via un questionnaire numérique.

Notre présentation relèvera les défis méthodologiques de l'usage de données authentiques dans une étude expérimentale. Plus précisément, nous aborderons les questions de la variabilité et de la représentativité de ce type de matériel, et par conséquent celle du contrôle des variables expérimentaux, ainsi que la question du rôle du contexte interactionnel dans le choix du matériel.

Références

[1] BARDOVI-HARLIG, K. & SALSBURY, T. The Organization of Turns in the Disagreements of L2 Learners : A Longitudinal Perspective. In D, Boxer & A. D. Cohen. (eds), *Studying Speaking to Inform Second Language Learning* (2004), 199-227. Clevedon : Multilingual Matters.

[2] BLOOMFIELD, A., WAYLAND, S. C., RHOADES, E., BLODGETT, A., LINCK, J. & ROSS, S. . *What makes listening difficult ? Factors affecting second language comprehension*. (2010). University of Maryland.

[3] GÜLICH, E. & MONDADA, L. « Analyse conversationnelle ». In Holtus, G., Metzeltin, M., Schmitt, C. (eds.). *Lexikon der Romanistischen Linguistik (LRL)* (2001), 196-250. Tübingen : Max Niemeyer Verlag.

[4] KENDRICK, K. H. Using Conversation Analysis in the Lab. *Research on Language and Social Interaction* 50, 1 (2017), 1-11.

[5] OURSEL, E. *Des interactions de service entre francophones natifs et non natifs, Analyse de la gestion de l'intercompréhension et perspectives didactiques*. (2013). Université de la Sorbonne nouvelle - Paris III.

[6] TAGUCHI, N. Instructed pragmatics at a glance : Where instructional studies were, are, and should be going. *Language Teaching* 48 (2015), 1-50.

Choix méthodologiques de recueil et d'analyse d'un corpus d'écrits scolaires corrigés : points forts et limites en lien avec l'outil de transcription

Sara MAZZIOTTI
4, rue des Irlandais (CLESTHIA) 75005 Paris, France
sara.mazziotti@studio.unibo.it

Cette recherche se situe au croisement entre les sciences du langage et les sciences de l'éducation est a pour objectif d'analyser un corpus de productions écrites recueillies dans deux écoles françaises et deux écoles italiennes à partir de la consigne "Que feras-tu quand tu seras grand ? Raconte une de tes journées". Ici, nous nous concentrerons sur les choix méthodologiques de recueil des textes en lien en particulier avec l'étape de transcription, afin d'en mettre en évidence leurs points forts et leurs limites. En effet, notre recherche a pour objectif de dessiner, d'une part, des possibles "profils d'enseignant-correcteur" français et italiens et d'autre part, d'évaluer la prise en compte de ces corrections de la part des élèves lors de la réécriture du brouillon. Cependant, nous essayerons également de confirmer l'hypothèse que les enseignants italiens corrigeraient moins l'aspect linguistique du texte par rapport aux enseignants français et qu'une surcharge de corrections dans le brouillon inhiberait la réélaboration du texte de la part de l'élève. Pour recueillir notre corpus scolaire, nous avons fait le choix d'assister aux deux différentes séances de rédaction d'une consigne commune dans huit classes de CE2 et de CM2 en France et en Italie. Nous avons demandé aux enseignants de corriger le brouillon selon leurs propres pratiques de correction et aux élèves d'en proposer une deuxième version à distance d'une semaine environ de l'écriture du premier jet.

L'outil d'analyse le plus important dont nous disposons est la transcription à partir du protocole utilisé par le groupe de recherche ECRISCOL (Claire Doquet) qui nous permet de signaler toutes les interventions de l'enseignant et l'évolution génétique du texte. Cependant, cet outil qui débute notre analyse ne peut que capturer les corrections que chaque enseignant a proposée dans ce seul brouillon à partir de la consigne prévue dans notre protocole, sans donner une vision plus complète de ces méthodes d'évaluation tout au long de l'année scolaire. De fait, la constitution d'un corpus impose des choix méthodologiques visant à le rendre cohérent et exploitable ensuite lors de l'analyse, mais aussi donc à limiter son extension.

Références

[4] Projet ECRISCOL (Doquet, Fleury) : http ://syled.univ-paris3.fr/ecriscol/CORPUS-TEST/

A computational approach to resolving the polysemy of postpositions in Korean

Seongmin Mun
MoDyCo (UMR 7114), CNRS, Paris Nanterre University
seongmin.mun@parisnanterre.fr

The current on-going project is for the resolution of polysemy involving the Korean postpositions. Korean is a Subject-Object-Verb language, which marks case with dedicated postpositions [1]. In this research project, we investigate the polysemy of postpositions in Korean under the framework of Construction Grammar [2]. A postposition is defined as a function word indicating grammatical information to which it is associated [1]. As a form-function pairing, a postposition can be polysemous in that one form delivers multiple functions [3]. An adverbial postposition -(u)lo, for instance, is either directional or instrumental, the two major functions of this particle [4] (1, 2).

(1) -(u)lo as directional ('(I) went to the road.')

도로-(으) 갔다.

tolo-(u)lo road-DIR

ka-ass-ta. go-PST-SE

(2) -(u)lo as instrumental ('(I) went by bicycle')

자전거-(으)로 갔다.

cacenke-(u)lo bicycle-INS

ka-ass-ta. use-PST-SE

We pose key questions as to what is polysemy of postpositions in Korean and how can computer identify the polysemy of the word? The meaning of a word in a sentence can be approximated by it's relation to the co-occurring words (dubbed the Distributional Hypothesis, [5]). It is thus assumed that we can identify the polysemy of a word based on information obtained from surrounding words and their network. This account has been implemented by way of NLP methods [6]. In this project, we use several NLP methods (such as SVD [7], PPMI SVD [8, 9], and SGNS [10, 11]) for the analysis of the Sejong corpus [12] which is made by large-scale corpus project in Korea in other to reveal the nature of polysemy involving postpositions in Korean. Currently, we are simultaneously progressing to develop the visualization and to make Gold standard from Sejong corpus. Gold standard includes target postpositions (-ey, -eyse, and -(u)lo) in the sentence and is designed to represent the functional semantic role of postposition (such as agent, experience, mental agent, companion, theme, location, direction, goal, final state, source, instrument, effector, criterion, purpose, content, etc.). With the visualization, we explore the information obtained from surrounding words and their network of a selected postposition. We will share the final system that can explore the distribution of polysemous postpositions and uses the distribution to automatically recognize the functional semantic role of postpositions in the upcoming conference.

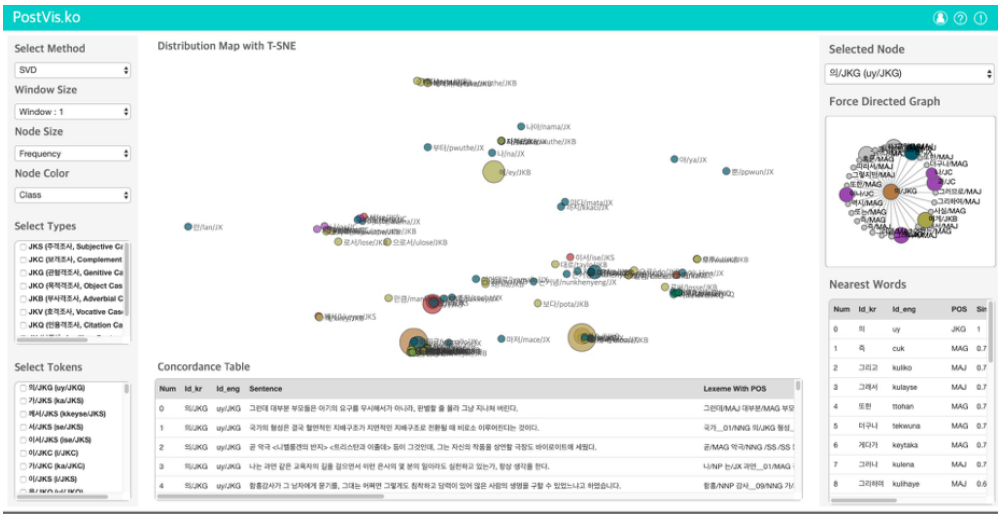


FIGURE 1 – PostVis.ko (URL : https ://seongmin-mun.github.io/VisualSystem/Major/PostVis/)

Références

[1] SOHN, H.-M. *The korean language*. Cambridge University Press, New York, NY. (1999).

[2] GOLDBERG, A. *Constructions : A construction grammar approach to argument structure*. Biblio-vault OAI Repository, the University of Chicago Press. (1995).

[3] GLYNN, D. AND ROBINSON, J. A. *Corpus methods for semantics. quantitative studies in polysemy and synonymy*. John Benjamins Publishing Company. (2014).

[4] CHOO, M. AND KWAK, H. Y. *Using korean*. Cambridge University Press, New York, NY. (2008).

[5] MCDONALD, S. AND RAMSCAR, M. Testing the distributional hypothesis : The influence of context on judgements of semantic similarity. In In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (2001), 611-6.

[6] HILPERT, M. Change in modal meanings. *Constructions and Frames* 8, 1 (2016), 66-85.

[7] ECKART, C. AND YOUNG, G. The approximation of one matrix by another of lower rank. *Psychometrika* 1, 3 (1936), 211-218.

[8] TURNEY, P. D. AND PANTEL, P. From frequency to meaning : Vector space models of semantics. *Journal of Artificial Intelligence Research* 37, 1 (2010), 41-188.
[9] BARONI, M. AND LENCI. A. Distributional memory : A general framework for corpus-based semantics. *Computational Linguistics* 36, 4 (2010), 673-721.

[10] MIKOLOV, T., CHEN, K., CORRADO, G. S., AND DEAN, J. Efficient estimation of word representations in vector space. In In *Proceedings of the International Conference on Learning Representations (ICLR)*. (2013).

[11] MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S., AND DEAN, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (2013), 3111-3119.

[12] HYO-PIL, S. The 21st sejong project : with a focus on selk(sejong electronic lexicon of korean) and the knc(korean national corpus). In *The 3rd International Joint Conference on Natural Language Processing*. (2008).

Quels choix méthodologiques pour la constitution d'un corpus d'ellipses erronées ?

Laura Noreskal
MoDyCo, UMR 7114, 200 Avenue de la République, 92000 Nanterre
l.noreskal@parisnanterre.fr

Notre travail porte sur les constructions elliptiques dans les copies des étudiants dans le cadre du projet **écri+** [1] qui vise le développement d'un dispositif national d'évaluation, de formation et de certification des compétences d'expression et de compréhension écrites en français. Après une première phase d'observation des erreurs dans les copies d'étudiants, il a été remarqué que les erreurs d'ellipses, bien que rares, constituaient une vraie zone « d'instabilité » lors de la rédaction, car plusieurs compétences linguistiques sont à maîtriser en même temps. En effet, les ellipses impliquent le respect de beaucoup de contraintes linguistiques assurant la stabilité de l'énoncé (Abeillé, 2008 ; Desmets, 2008, Bîlbîie, 2011). Le non-respect de ces contraintes entraîne alors une incohérence de l'énoncé (Marie aime le chocolat chaud et Pierre **[aime]** le thé. / *Anne mange du gâteau et Jean **[mange]** une part.). De ce fait, il nous a semblé important de porter nos recherches sur ce phénomène, particulièrement révélateur des compétences rédactionnelles des étudiants ; ainsi notre contribution au dispositif national d'évaluation sera de créer un outil capable de détecter et de corriger automatiquement les ellipses erronées dans les textes brefs. La première étape de notre recherche a été de constituer un corpus d'ellipses erronées issues de rédactions d'étudiants afin d'identifier les erreurs d'ellipses dans ces productions et de tester plusieurs techniques de TAL pour la détection et la résolution automatique des ellipses. Lors de cette communication, notre objectif sera de décrire la méthodologie adoptée pour la constitution de notre corpus. La première partie sera consacrée à l'influence des choix méthodologiques sur notre recueil de données. Nous expliquerons le choix des productions étudiantes pour constituer notre corpus et nous présenterons également les différents types de productions que nous avons pu collecter. Dans cette partie, nous développerons alors nos premières observations sur les différences d'apparition d'ellipses erronées entre les rédactions spontanées (partiels et exercices faits en classe) et les rédactions préparées (devoirs maison, rapports de stage et mémoires) et nous expliquerons comment ces observations ont influencé nos derniers résultats. En deuxième lieu, nous présenterons le choix effectué concernant les métadonnées retenues, comme le type de rédaction, le niveau et le domaine d'études de l'étudiant en expliquant les points fort et les limites de nos données. Ce choix étant particulièrement important pour la réalisation d'analyses sociolinguistiques sur le phénomène de l'ellipse. Pour finir, nous présenterons les perspectives du travail qui concernent les méthodes de TAL que nous voulons appliquer sur le corpus constitué afin de détecter automatiquement les ellipses erronées.

1. Numéro ANR : ANR17NCUN0015

Références

[1] ABEILLÉ A. & MOURET F. Quelques contraintes sur les coordinations elliptiques en français. *Revue de sémantique et de pragmatique* 24, 89 (2008), 177-206.

[2] BILBIIE G. *Grammaire des constructions elliptiques. Une étude comparative des phrases sans verbe en roumain et en français*, Thèse de Doctorat, Université Paris 7. (2011).

[3] DESMETS M. Ellipses dans les constructions comparatives en comme, *Linx. Revue des linguistes de l'université Paris X Nanterre* 58 (2008), 47-74.

La chanson chrétienne : pratique, fonctionnalités des langues et expression des dynamiques plurilingues au Cameroun. Une articulation entre données écrites et données orales

Michel Narcisse Ntedondjeu
Université Sorbonne Nouvelle-Paris 3 & Université de Buea, Cameroun
ntedondjeu2004@yahoo.fr

« Contrairement aux autres types de chansons, la chanson religieuse chrétienne a une origine biblique dans la mesure où les textes des chants liturgiques s'inspirent en grande partie de l'Ancien et du Nouveau Testament en s'enrichissant toute fois des rencontres faites au cours des siècles avec des personnages ou des événements qui ont marqué l'Eglise » [1]. Leur étude peut présenter de nombreux intérêts pour la sociolinguistique dans la mesure où la chanson religieuse reflète tout autant la dynamique linguistique que sociale d'une communauté. En effet, l'enracinement des communautés concernées par cette recherche dans leur milieu culturel leur a permis de créer des modèles comportementaux caractérisant les communautés de pratiques plurilingues. Ainsi la chanson religieuse n'est pas en marge de cette réalité sociolinguistique plurielle dans laquelle elle est exécutée. Cette recherche est consacrée à ce genre discursif particulier propre aux cérémonies religieuses. Le choix de ce genre discursif parmi les autres en présence est motivé par sa richesse du point de vue des langues exploitées pour son exécution, mais également, par les modalités d'exécution de la chanson religieuse lors des offices qui montrent que le choix des chansons et des langues n'est pas neutre. Il sera question d'étudier les chansons qui accompagnent les offices dans trois communautés chrétiennes situées dans une ville moyenne du Cameroun : Buea. À partir d'un corpus pluriel constitué de textes de chansons religieuses (programmes de messe, observations participantes, entretiens menés auprès des acteurs de la communication religieuse), nous montrerons, en articulant données écrites (chansons) et orale (entretiens) comment le chant participe à rythmer les cérémonies religieuses, à les structurer et produire de l'unité dans la diversité. Notre approche qui se veut ethnographique va permettre l'étude des chansons dans leur thématique, leur emploi et dans certaines de leurs caractéristiques langagières ; ce qui va aider à comprendre leur rôle au sein des communautés et au cœur des offices. En s'intéressant à la façon dont ces chansons participent à l'accomplissement des rites et à leurs significations, nous allons également montrer comment le chant religieux favorise la mise en œuvre du plurilinguisme et comment les acteurs religieux se situent par rapport à la diversité linguistique et l'envisagent. Cette recherche considère la communauté religieuse comme une communauté de pratique [2] et donc la chanson chrétienne exécutée in situ comme une de ces pratiques à travers lesquelles la participation et l'expérience des membres de la communauté se figent. Par notre analyse textuelle [3], il s'agira d'étudier la dynamique des fonctionnements discursifs des chants et des entretiens aux travers de leurs modalités d'énoncé, des catégories référentielles, lexicales, syntaxiques et énonciatives construites en articulation avec la situation de communication, le cotexte et le contexte.

Références

[1] DAMON-GUILLOT, A. « La mémoire dans la musique liturgique de l'Église chrétienne orthodoxe unifiée d'Éthiopie. à travers la performance, l'écriture et la rencontre ». *Cahiers d'ethnomusicologie. Anciennement Cahiers de musiques traditionnelles* 22 (2009), 187-201.

[2] WENGER, E. *La théorie des communautés de pratiques. Apprentissage, sens et identité* (2005), Sainte Foy-Presse de l'Université de Laval.

[3] ADAM, J-M. *La linguistique textuelle. Introduction à l'analyse textuelle des discours* (2005 [2008]). Paris, Armand Colin, collection "Cursus".

Une étude empirique des périphrases verbales progressives en roumain et en français

Beatrice Pahontu
Faculté de Langues et Littératures Etrangères, Univ. de Bucarest
Pahontubeatrice.pahontu@yahoo.com

Cette étude vise la description des périphrases verbales progressives dans deux langues romanes : le français et le roumain, en particulier les constructions avec être en train de en français et *a fi pe cale să/de* 'être en train de' et *a fi în curs de/să* 'être en cours de' en roumain, considérées comme des structures marginales (avec une fréquence réduite dans ces deux langues, cf. Bertinetto 2000). Si pour le français notre étude contribue à tester en corpus les hypothèses avancées dans la littérature (en particulier, la préférence pour les sujets animés cf. Mortier 2005, l'incompatibilité avec les temps perfectifs cf. Squartini 1998, la fréquence réduite avec des prédicats sémantiques de type achèvement et l'incompatibilité avec les états cf. Mortier 2005, etc.), pour le roumain, cela constitue la première analyse empirique des deux périphrases progressives.

Du côté du français, nous avons envisagé une perspective diachronique et synchronique pour observer la dynamique de la périphrase *être en train de*, dont la fréquence semble être plus élevée dans le passé (Schøsler 2007). Une telle analyse permet de mieux saisir le passage des emplois modaux de cette construction vers les emplois progressifs, évolution propre au processus de grammaticalisation : le sens progressif est rencontré sporadiquement au XVIIIe siècle, mais c'est vers la deuxième moitié du XIXe siècle qu'il devient dominant (Gougenheim 1929, Schøsler 2007). Pour cette première étape de notre travail, nous avons donc pris en compte des données relevant du XIXe siècle et du XXe siècle, extraites de la base de données Frantext. Pour le roumain, nous nous sommes servis d'un corpus qui regroupe des textes littéraires et scientifiques de plusieurs domaines (médecine, droit, histoire, biologie, agronomie, critique et théorie littéraire, philosophie), relevant du XVIe siècle jusqu'à la deuxième moitié du XXe siècle et du corpus CoRoLa (*Corpus of Contemporary Romanian Language*). Les périphrases du roumain étant relativement récentes (fin du XIXe siècle - début du XXe siècle), nous avons adopté une perspective synchronique.

Nous présenterons la grille d'annotation établie en fonction des critères morpho-syntaxiques et sémantiques et les restrictions de sélection de chaque périphrase, comme révélé par notre corpus.

Références

[1] BERTINETTO, P. M. The progressive in Romance, as compared with English. In Östen Dahl (éd.), *Tense and Aspect in the language of Europe* (2000), 559–664. Berlin/New York : Mouton de Gruyter.

[2] GOUGENHEIM, G. *Étude sur les périphrases verbales de la langue française*. Paris : Les Belles Lettres. (1929).

[3] MORTIER, L. Les Périphrases aspectuelles « progressives » en français et en néerlandais. Présentation et voies de grammaticalisation ». In Hava Bat-Zeev Shyldkrot et Nicole Le Querler (éds.), *Les Périphrases verbales* (2005), 83-102. Amsterdam / Netherlands : John Benjamins Publishing.

[4] SCHØSLER, L. Grammaticalisation et dégrammaticalisation : Étude des constructions progressives en français du type Pierre va / vient / est chantant. In Emmanuelle Labeau, Carl Vetters et Patrick Caudal (éds.), *Sémantique et diachronie du système verbal français* (2007), 91-119. Amsterdam/Netherlands : Rodopi.

[5] SQUARTINI, M. *Verbal Periphrases in Romance. Aspect, Actionality, and Grammaticalization*. Berlin/New York : Mouton de Gruyter. (1998).

Représentations identitaires de la radicalité djihadiste sur Facebook par l'analyse outillée de corpus : de l'enquête de terrain à l'analyse de données sociales

Laurène Renaut
Université de Cergy-Pontoise
laurene.renaut@orange.fr

Cette communication qui s'inscrit au croisement de plusieurs courants des Sciences du langage (analyse du discours, sociolinguistique et linguistique de corpus) se propose d'interroger, dans le contexte spécifique de la radicalisation djihadiste en ligne donc de la djihadosphère, les méthodes pour exploiter des données issues de réseaux ou médias sociaux, et cela à différentes étapes du processus de recherche : constitution d'un corpus natif en ligne, collecte des données, caractérisation de leur contenu et analyse de leur dynamique. Notre recherche s'appuie en effet sur l'étude des données publiques de profils radicalisés sur Facebook ; donc sur un corpus numérique anonymisé pour des raisons de sécurité comme de confidentialité. Une grille de critères a été préétablie dans le cadre de ce travail afin de définir l'ensemble des éléments caractéristiques attestant d'une forme de radicalisation sur ce réseau social, et permettant ainsi de circonscrire le terrain analysé. Face aux phénomènes émergents de cyberviolence et de propagation d'idéologies extrêmes sur le web, notre étude propose plus précisément de questionner les représentations identitaires de cette « communauté communicante radicale » qui semble mettre en œuvre une hétérotopie de crise [1]. Dessinant en effet un espace réservé aux personnes qui rejettent les règles de la société dans laquelle ils vivent, cette hétérotopie numérique renvoie ainsi à la création d'un nouveau lieu (ou tiers-lieu) qui fait coexister deux mondes dans un même espace : le monde réel (l'ici et maintenant dont l'idéologie dominante est critiquée, déconstruite et combattue) et l'ailleurs (un autre monde, fantasmé et renvoyant à une utopie révolutionnaire qui serait ici le Califat). Ce « contre-emplacement » qui relève ainsi d'un monde virtuel mais dans un espace bien réel (ancré dans l'interface de Facebook) s'apparente à un lieu carnavalesque [2] dans une logique de spectacle et d'inversion des normes. Déplaçant ainsi les participants de la sphère officielle vers la sphère non-officielle, d'où ils peuvent expérimenter la transgression de la norme tout en étant encadrés par la sphère officielle (censure de Facebook, surveillance des services de renseignements…) cet espace est celui d'acteurs antisystème qui s'exprime paradoxalement dans le système. C'est dans ce contexte que nous interrogerons ainsi les évolutions des stratégies sémio-discursives déployées pour se dire « djihadiste » sur les réseaux sociaux entre 2015 et 2019 ? Nous avons d'abord choisi une approche sans interaction directe avec les utilisateurs et par laquelle les usagers sont observés à distance. Prenant appui sur les travaux de [3] nous avons mené nos recherches dans la perspective d'une sémiotique appliquée de l'identité en ligne. Par ailleurs, nous considérons le corpus comme un « terrain » [4], lequel a été préparé en amont par l'observation et l'établissement de questions de recherche préalables afin d'être questionné avec les outils offerts par une analyse sémiolinguistique.

Références

[1] FOUCAULT, M. *Le Corps utopique*, suivi de *Les Hétérotopies* (avec une postface de Daniel Defert). (1966).

[2] BAKHTINE, M. Esthétique et théorie du roman. Trans. Daria Olivier. (1978). Paris : Gallimard.

[3] PERAYA, D. « Les changements induits par les technologies », in *Actes du Colloque CETSIS-EEA 99*, Montpellier, université de Montpellier II, Cépaduès (1999), 185-188.

[4] LONGHI, J. *Du discours comme champ au corpus comme terrain*. Paris : L'Harmattan. (2018).

Établir des corpus linguistiques bilingues et comparables en FLE

Liu YANG
Université Paris-Sorbonne
minot8627@gmail.com

Depuis la fin du XXe siècle, surtout après la publication du CECRL (2001), la didactique du FLE attache de l'importance à la compétence (inter)culturelle des apprenants, en prenant en compte leur propre culture éducative (Spaëth, 2004). Ainsi, l'introduction des corpus linguistiques bilingues et comparables semble nécessaire et favorable. Notamment pour des langues éloignées, comme le français et le chinois, il est important d'établir une interaction entre les unités et les faits linguistiques de ces deux langues. La morphologie et la syntaxe spécifiques du chinois révèlent un grand décalage avec les autres langues occidentales. Un événement, une graphie et un signifiant constituent un caractère chinois, ayant des sens riches et des usages variables (figure 1) (Ryjik, 1983). Mais comme le chinois est une langue où la pragmatique est dominante et que, le chinois manque de marques morphologiques et syntaxiques, l'usage des mots chinois dépend fortement du contexte (Lu, 2015 :23). Selon Giroux, c'est sur la langue maternelle que l'apprenant « prend généralement appui pour aborder l'apprentissage d'une autre langue » (2016 :64). De plus, l'apprentissage d'une langue étrangère est une procédure « d'intégration » (Biiclé, 2009 : 36). Comment alors intégrer le français dans le système langagier des apprenants sinophones ? Il nous faut pour cela recourir à des corpus linguistiques bilingues et comparables, sachant qu'un corpus comparable désigne « des données collectées à l'aide d'un même type d'échantillon et d'une représentativité similaire. » (Hennecke, 2018). La comparaison s'effectue ici via des œuvres littéraires et leurs traductions, donc, il s'agit de comprendre « comparables » comme « traduits ». Dans cette optique, j'ai importé une œuvre littéraire française et sa version chinoise sur un logiciel, AntConc, pouvant traiter parfaitement le chinois et le français. À travers ces deux corpus en français et en chinois construits, on peut consulter facilement les fréquences, les occurrences et les contextes des mots, des locutions et des unités linguistiques, en comparant leurs sens et leurs fonctions dans ces deux langues. L'apprenant peut donc déduire l'emploi des unités linguistiques compliquées, en y établissant un schéma cognitif (figure 2). Le corpus choisi est Boule de suif de Maupassant et sa traduction, étant donné qu'il est très connu des Chinois. L'objectif pédagogique est de faire comprendre les articles définis et indéfinis à une vingtaine d'apprenants chinois de niveau A2. Les résultats montrent que les enjeux des corpus bilingues ne se limitent pas à l'acquisition des connaissances linguistiques, mais concernent aussi une communication interculturelle chez les apprenants sinophones.

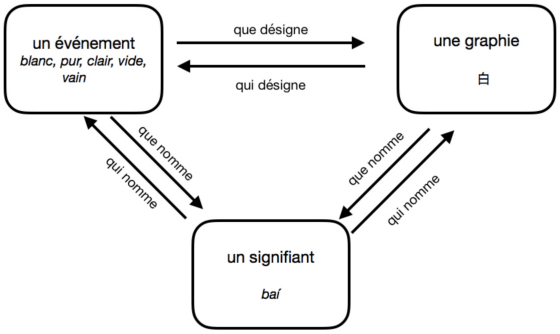


FIGURE 1 – Schéma de trois secteurs de caractère chinois

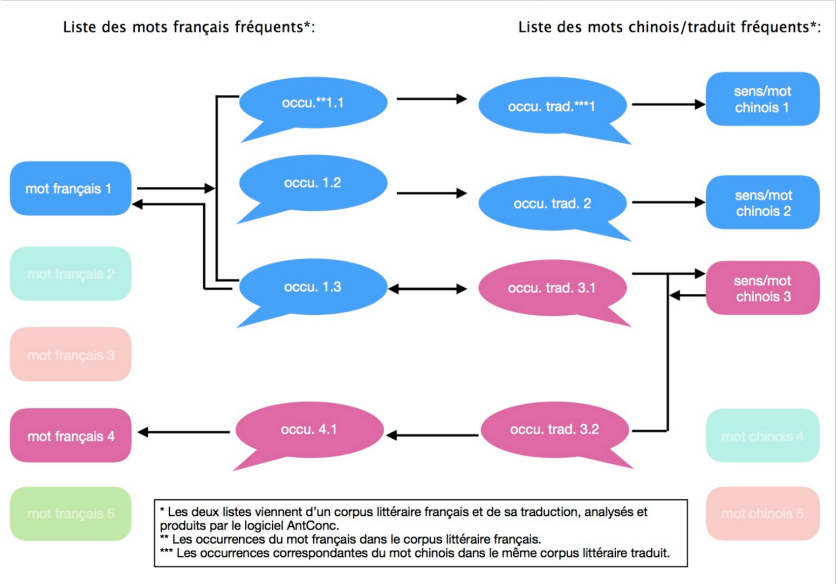


FIGURE 2 – Schéma cognitif

Références

[1] BIICHLÉ L. « Le plurilinguisme c'est l'intégration ». *Savoirs et formation* 73 (2009), 32-35. HAL.

[2] Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer. Conseil de l'Europe, Division des Langues Vivantes, Paris, Didier. (2001).

[3] GIROUX L. « La place et le(s) rôle(s) de la langue maternelle des apprenants en cours de langue étrangère ». *Synergies France* 10 (2016). 55-68.

[4] HENNECKE, I. « Petits corpus oraux bilingues et plurilingues–enjeux théoriques et méthodologiques ». *Corpus* 18 (2018).

[5] LU J. « Hànyǔ 'tèdiǎn'zhīwǒjiàn (les caractéristiques du chinois à mes yeux) ». *Journal of Chiniese studies* (2015), 15-26. Xiamen University.

[6] RYJIK K. *L'idiot chinois : Initiation élémentaire à la lecture intelligible des caractères chinois*. (1983). Paris, Payot.

[7] SPAËTH V. « Le concept de 'langue-culture' et ses enjeux contemporains dans l'enseignement/apprentissage des langues » *L'enseignement de l'arabe en Israël et en France ; l'enseignement de l'hébreu dans le monde arabe : des regards croisés* (2014), 1-17.