

The logo for CRISCO, consisting of the word 'CRISCO' in a blue, sans-serif font with a light blue shadow effect. Below it, the words 'Langues | Signification | Contexte' are written in a smaller, grey font.

CRISCO
Langues | Signification | Contexte

The logo for inalco CREE, featuring the word 'inalco' in a blue, lowercase sans-serif font, followed by 'CREE' in a larger, bold, blue, uppercase sans-serif font. Below this, the text 'Centre de recherche Europes-Eurasie' is written in a smaller, grey font.

inalco
CREE
Centre de recherche
Europes-Eurasie

The logo for CEL, featuring the text 'UNIVERSITÉ JEAN MOULIN LYON 3' in a small, grey font around the top edge of a white circle. Inside the circle, there is a blue stylized 'C' and 'E' logo. To the right of the circle, the letters 'CEL' are written in a large, bold, blue, sans-serif font. Below this, 'Centre d'Études Linguistiques' is written in a smaller, grey font.

UNIVERSITÉ JEAN MOULIN LYON 3
CEL
Centre d'Études
Linguistiques

The logo for inalco, featuring the word 'inalco' in a blue, lowercase sans-serif font. Above it, the word 'national' is written in a smaller, grey font, and above that, the characters '↑**' and '↑**' are visible. To the right of 'inalco', the words 'национален', 'עבר', '文化', and 'ثقافة' are written in their respective languages. Below 'inalco', the text 'Institut national des langues et civilisations orientales' is written in a small, grey font.

↑**↑**
национален עבר 文化 ثقافة
inalco
Institut national
des langues
et civilisations orientales

La linguistique de l'oral spontané à travers les langues

**création, annotation et analyse de
corpus, segmentation du discours**

**Colloque
international**

**23 et 24 mai
2024**

**CREE INALCO Paris
CEL Université Jean Moulin Lyon 3
CRISCO Université Caen Normandie**

SOMMAIRE



Programme.....	3
Conférences plénières.....	5
Virginie ANDRÉ, Annoter un corpus d'interactions pour faciliter l'enseignement et l'apprentissage du français langue étrangère.....	9
Sara BADIA-CLIMENT, Guidelines for the construction of a segmented oral corpus based on the Val.Es.Co. 3.0 corpus.....	11
Fatma BEN BARKA-MESSAOUDI, Rayan ZIANE, Meriem BERAÏK, Pour la segmentation automatique de l'arabe parlé : l'exemple de l'arabe tunisien.....	13
Mireille BILGER, Paul CAPPEAU, De l'intérêt de corpus diversifiés.....	16
Mylène BLASCO, Paul CAPPEAU, La parole des médecins : un oral sous l'influence de l'écrit ?.....	18
Outi DUVALLON, Il-Il YATZIV-MALIBERT, L'analyse syntaxique en « grilles » des textes oraux : illustrations avec des données de finnois et d'hébreu moderne.....	20
Kirill GANZHA, Paramétrage et analyse d'un corpus oral modélisé en LV2 russe : étude à des fins pédagogiques.....	22
Marina HAAN, Segmentation du discours de débat en anglais langue étrangère. Comment identifier l'unité spontanée dans l'oral préparé ?.....	24
Karolina KOPCZYŃSKA, Donc euh voilà donc je manque de vocabulaire. Pauses remplies et répétitions en tant que stratégies employées pour organiser le discours spontané à l'exemple du mini- corpus de français parlé en Pologne.....	26
Loïc LIÉGEOIS, Christophe BENZITOUN, Christine DA SILVA-GENEST, Caroline MASSON, Christophe PARISSÉ, Méthodologie(s) de segmentation des transcriptions de l'oral en vue de la création d'une mesure d'évaluation des productions verbales enfantines.....	28
Karine MARTEL, Interactions dans le contexte du polyhandicap.....	31
Philippe MARTIN, Annotation prosodique dans l'oral spontané.....	33
Erwanne MAS, How effective are automatic annotation tools for investigating vocalic change in Australian English? An analysis of conversational speech.....	35
Samira MOUKRIM, La linguistique de l'oral à travers les langues. Sur les problèmes de transcription des langues parlées peu dotées.....	37
Adrià PARDO LLIBRER, On the two realms of spoken Spanish: a study based on discourse markers positions...	41
Liubov PATRUKHINA, Transcrire ou ne pas transcrire, telle est la question : le choix de transcription de l'oral dans le cadre d'un projet européen.....	43
Laurent PRÉVOT, Roxane BERTRAND, Unités Discursives, Chunks Syntaxiques et Syntagmes Accentuels en Français Conversationnel.....	45
Thierry RUCHOT, Intralocutionary and interlocutionary coherence.....	48
Valentina SACCONI, Lorenzo GREGORI, Shuai LUO, Segmenting Spontaneous Spoken Language: An Exploration of Prosody and Perceptual Strategies in Italian and Chinese.....	50
Elena VLADIMIRSKA, Les indices mimique-gestuels comme marqueurs des opérations énonciatives.....	53

PROGRAMME

Jeudi 23 mai 2024 – 65 rue des Grand Moulins, Paris 75013		
9h00 – 9h30	Accueil des participants et mot de bienvenue	
CONFÉRENCE PLÉNIÈRE 1		
9h30 – 10h30	Bernd Heine On the role of interactives in the organization of spoken discourse	
10h30 – 10h50	Pause-café	
SESSION 1	10h50 – 11h10	Thierry Ruchot Intralocutionary and interlocutionary coherence
	11h10 – 11h30	Martel Karine Interactions dans le contexte du polyhandicap
	11h30 – 11h45	Discussion
	11h45 – 12h05	Mireille Bilger, Paul Cappeau De l'intérêt de corpus diversifiés
	12h05 – 12h25	Samira Moukrim La linguistique de l'oral à travers les langues Sur les problèmes de transcription des langues parlées peu dotées
	12h25 – 12h40	Discussion
12h40 – 14h00	Pause déjeuner	
SESSION 2	14h00 – 14h20	Mylène Blasco, Paul Cappeau La parole des médecins : un oral sous l'influence de l'écrit ?
	14h20 – 14h40	Liubov Patrukhina Transcrire ou ne pas transcrire, telle est la question : le choix de transcription de l'oral dans le cadre d'un projet européen
	14h40 – 15h00	Loïc Liégeois, Christine da Silva-Genest, Christophe Benzitoun, Caroline Masson et Christophe Parisse Méthodologie(s) de segmentation des transcriptions de l'oral en vue de la création d'une mesure d'évaluation des productions verbales enfantines
	15h00 – 15h20	Sara Badia Climent Guidelines for the construction of a segmented oral corpus based on the Val.Es.Co. 3.0 corpus
	15h20 – 15h50	Discussion
15h50 – 16h10	Pause-café	
SESSION 3	16h10 – 16h30	Valentina Saccone, Lorenzo Gregori, Shuai Luo Segmenting Spontaneous Spoken Language: An Exploration of Prosody and Perceptual Strategies in Italian and Chinese
	16h30 – 16h50	Fatma Ben Barka-Messaoudi, Rayan Ziane, Meriem Beraik Pour la segmentation automatique de l'arabe parlé : l'exemple de l'arabe tunisien
	16h50 – 17h10	Marina Haan Segmentation du discours de débat en anglais langue étrangère. Comment identifier l'unité spontanée dans l'oral préparé ?
	17h10 – 17h30	Discussion

Vendredi 24 mai 2024 – 2 rue de Lille, Paris 75007

CONFÉRENCE PLÉNIÈRE 2

	9h30 - 10h30	Salvador Pons Bordería If not Syntax, then what? Insights from the Val.Es.Co. model of discourse segmentation
	10h30 - 10h50	Pause-café
SESSION 1	10h50 - 11h10	Adrià Pardo Llibrer On the two realms of spoken Spanish: a study based on discourse markers positions
	11h10 - 11h30	Virginie André Annoter un corpus d'interactions pour faciliter l'enseignement et l'apprentissage du français langue étrangère
	11h30 - 11h50	Erwanne Mas How effective are automatic annotation tools for investigating vocalic change in Australian English? An analysis of conversational speech
	11h50 - 12h05	Discussion
SESSION 2	12h05 - 12h25	Philippe Martin Annotation prosodique dans l'oral spontané
	12h25 - 12h45	Elena Vladimirska Les indices mimico-gestuels comme marqueurs des opérations énonciatives
	12h45 - 13h00	Discussion
	13h00 - 14h00	Pause déjeuner
CONFÉRENCE PLÉNIÈRE 3		
	14h00 - 15h00	Gunther Kaltenböck On the recent emergence of a new discourse marker
SESSION 3	15h00 - 15h20	Outi Duvallon, Il-Il Yatziv-Malibert L'analyse syntaxique en « grilles » des textes oraux : illustrations avec des données de finnois et d'hébreu moderne
	15h20 - 15h40	Laurent Prévot & Roxane Bertrand Unités Discursives, Chunks Syntaxiques et Syntagmes Accentuels en Français Conversationnel
	15h40 - 16h00	Discussion
	16h00 - 16h20	Pause-café
SESSION 4	16h20 - 16h40	Kirill Ganzha Paramétrage et analyse d'un corpus oral modélisé en LV2 russe : étude à des fins pédagogiques
	16h40 - 17h00	Karolina Kopczyńska <i>Donc euh voilà donc je manque de vocabulaire.</i> Pauses remplies et répétitions en tant que stratégies employées pour organiser le discours spontané à l'exemple du mini-corpus de français parlé en Pologne
	17h00 - 17h20	Discussion et mot de clôture

ON THE ROLE OF INTERACTIVES IN THE ORGANIZATION OF SPOKEN DISCOURSE



Bernd HEINE

University of Cologne

bernd.heine@uni-koeln.de

The paper to be presented is based on the model of interactive grammar, resting on the extra-clausal expressions of linguistic discourse, are distinguished, namely attention directives, discourse markers, evaluatives, ideophones, interjections, response elicitive for organizing parts of spoken discourse. It is argued that the concept of valency can be of help for understanding some relations of dependency obtaining between segments of texts.

References

HEINE, Bernd. 2023. *The Grammar of Interactives*. Oxford: Oxford University Press.

Biography

Bernd Heine is Emeritus Professor at the Institute of African Studies and Egyptology, University of Cologne. He has held visiting professorships in Europe, USA, S. Korea, Japan, China, Australia, Brazil, Mexico, Australia, Kenya, and South Africa. His main research areas are endangered languages in Africa, grammaticalization theory, and discourse grammar. He has been a fellow of the Center for Advanced Study in the Behavioral Sciences, Stanford, USA (1999-2000), the Netherlands Institute for Advanced Study (NIAS), Wassenaar (2005-2006), Distinguished World-Class Scholar, Ministry of Education, South Korea (2008), and Tokyo University of Foreign Studies (2008-9). In 2009, he received the Life Time Achievement Award of the Evolutionary Linguistics Association.

His book publications include *Possession: Cognitive Sources, Forces, and Grammaticalization* (CUP, 1997); *Cognitive Foundations of Grammar* (OUP, 1997); *Language Contact and Grammatical Change* (CUP, 2005), *The Genesis of Grammar: A Reconstruction* (OUP, 2007); *World Lexicon of Grammaticalization* (CUP, 2019, with co-authors), and *The Grammar of Interactives* (OUP, 2023).

IF NOT SYNTAX, THEN WHAT? INSIGHTS FROM THE VAL.ES.CO. MODEL OF DISCOURSE SEGMENTATION



Salvador PONS BORDERÍA

University of Valencia – Val.Es.Co

salvador.pons@uv.es

Over the past fifty years, research stemming from foundational inquiries into the structure of spoken language (del Val 1966; Narbona 1979; Roulet et al. 1985; Blanche-Benveniste and Jeanjean 1987) has propelled significant advancements in Romance linguistics. Pragmatic contributions have enriched a unified theoretical framework, now undergirded by robust spoken language corpora such as PILEI, Val.Es.Co., and PRESEEA, offering invaluable resources to the scholarly community.

Progressing past an initial phase characterized by exploratory analysis and detailed focus, the field is poised to enter a new era of synthesis and empirical validation. The current trajectory emphasizes the systematization and statistical evaluation of linguistic data, facilitated by a critical assessment of competing theoretical perspectives. The ultimate goal is a comprehensive understanding of the structure of spoken language. This presentation aims to delineate the essential criteria for these theoretical constructs and to prognosticate their potential utility in ongoing linguistic inquiry.

Biography

Salvador Pons Bordería is Professor of Spanish Linguistics at the University of Valencia, Spain. Founder member of the Val.Es.Co. Research Group (Valencia Español Coloquial), prof. Pons has published more than one hundred articles and other contributions in top ranked journals and printers like Journal of Pragmatics, Revista de Filología Española, John Benjamins and Oxford University Press. He has been keynote speaker in more than 20 universities in Europe, Asia and America. Since 2016, he is also a member of the IPrA Advisory Board.

His current research covers the fields of corpus linguistics (www.valesco.es) and the lexicography of discourse markers (www.dpde.es); the structure of spontaneous conversations; the constructionalization processes of DMs, and the diachronic study of the XXth century.

JUST SO YOU KNOW: ON THE RECENT EMERGENCE OF A NEW DISCOURSE MARKER



Gunther KALTENBÖCK

University of Cologne

gunther.kaltenboeck@uni-graz.at

This presentation identifies just so as a newly emerging purpose subordinator. Using data from the Corpus of Contemporary American English (Davies 2008-) and the Corpus of Historical American English (Davies 2010), it traces its development and steady increase in frequency from its first attestation in the mid-19th century to the present day. Just so is shown to represent a case of semantic specialization where the purpose meaning wins out over the conditional meaning, thus filling the niche of an informal purpose subordinator and providing an alternative to its multifunctional and semantically ambiguous competitors so that and so (Kaltenböck & ten Wolde, 2022). With increasing grammaticalization the just so purpose subordinator also exhibits signs of intersubjectification, being coopted for syntactically independent, interpersonal uses (e.g. Just so we're clear) and culminating in the emergence of a new discourse marker in the form of just so you know in the late 20th and early 21st century (Kaltenböck et al. 2011, Heine et al. 2017, 2021). To account for the emergence of purpose just so, a constructional network approach is adopted, which considers the network links to other purpose subordinators, notably so that and so (Diessel 2019).

References

- DAVIES, Mark. 2008. The corpus of contemporary American English (COCA). Available online at <https://www.english-corpora.org/coca/>.
- DAVIES, Mark. 2010. The corpus of historical American English (COHA). Available online at <https://www.english-corpora.org/coha/>.
- DIESSEL, Holger. 2019. The grammar network. How linguistic structure is shaped by language use. Cambridge: Cambridge University Press.
- HEINE, Bernd, Gunther KALTENBÖCK, Tania KUTEVA & Haiping LONG. 2021. The rise of discourse markers. Cambridge: Cambridge University Press.
- HEINE, Bernd, Gunther KALTENBÖCK, Tania KUTEVA, & Haiping LONG. 2017. Cooptation as a discourse strategy. *Linguistics* 55(4). 1–43.
- KALTENBÖCK, Gunther, Bernd HEINE & Tania KUTEVA. 2011. On thetical grammar. *Studies in Language* 35(4). 848–893.
- KALTENBÖCK, Gunther & Elnora TEN WOLDE. 2022. A just so story: On the recent emergence of the purpose subordinator just so. *English Language & Linguistics* 26(4). 889–915.

Biography

Gunther Kaltenböck is Professor of English Linguistics at the University of Graz, Austria. His research interests lie in the areas of cognitive-functional grammar, corpus linguistics, pragmatics, variation and change. Apart from numerous book chapters and contributions to international journals, his publications include a co-authored book on *The rise of discourse markers* (2021, Cambridge University Press), a monograph on *It-extraposition and non-extraposition in English* (2004, Braumüller) and several co-edited volumes, such as *Outside the clause* (2016, Benjamins), *Insubordination: Theoretical and empirical issues* (2019, de Gruyter) and *Grammar and cognition: Dualistic models of language structure and language processing* (2020, Benjamins), and *Special Issues, such as Investigating stance constructions in English: Synchrony and diachrony with Language Sciences* (2020) and *Fragments: Construction and reconstruction with English Language and Linguistics* (2024 forthc.).

ANNOTER UN CORPUS D'INTERACTIONS POUR FACILITER L'ENSEIGNEMENT ET L'APPRENTISSAGE DU FRANÇAIS LANGUE ÉTRANGÈRE



Virginie ANDRÉ

Université de Lorraine et CNRS – ATILF

virginie.andre@univ-lorraine.fr

Dans le cadre du projet FLEURON (français langue étrangère universitaire – ressources et outils numériques), nous avons élaboré un dispositif d'apprentissage (<https://fleuron.atilf.fr/>) constitué d'un corpus multimodal d'interactions authentiques interrogeable par un concordancier (André, 2016). Ce corpus est composé d'une part d'interactions sollicitées spontanées et d'une part d'interactions non sollicitées spontanées, filmées en situation écologique. Le dispositif FLEURON étant dédié aux étudiants étrangers souhaitant faire ou faisant un séjour universitaire en France, les situations d'interaction tentent d'être représentatives de celles susceptibles d'être vécues par un étudiant. Elles concernent des interactions relatives à la vie universitaire (scolarité, services du CROUS, au sein d'associations, etc.) et des interactions quotidiennes (dans des commerces, des lieux culturels, des services de santé, entre amis, etc.).

Le dispositif d'apprentissage permet aux apprenants d'observer et d'analyser le déroulement des interactions qui sont transcrites et sous-titrées ainsi que d'interroger les transcriptions avec le concordancier associé et de mettre en œuvre la démarche du data-driven learning (Johns, 1991 ; André & Ciekanski, 2018 ; Boulton & Vyatkina, 2021). Le dispositif peut être utilisé dans différentes configurations, en autonomie ou en salle de classe avec plus ou moins de guidance. Les activités cognitives que les apprenants réalisent lors de l'utilisation du dispositif, notamment en observant les données, en les analysant et en induisant des règles de fonctionnement, font émerger des moments d'apprentissage de la langue (André, 2023).

Afin d'aider les enseignants et les apprenants à observer et à saisir le fonctionnement des interactions, nous créons des fiches descriptives qui présentent différentes couches d'annotation : les séquences (ouverture, clôture, par exemple), les activités langagières (saluer, demander, préciser, expliquer, interrompre, etc.), les micro-phénomènes spécifiques du français parlé en interaction et la multimodalité (Traverso, 2016). Ces annotations sont donc destinées à des personnes qui ne sont pas formées à l'analyse des interactions mais qui ont besoin de comprendre les pratiques langagières utilisées par les locuteurs. Nous annotons tout particulièrement les éléments intéressants pour des apprenants et des enseignants de FLE et qui sont généralement absents des manuels de langue et des grammaires traditionnelles (Surcouf, 2022).

Nous présenterons les difficultés rencontrées pour proposer différentes couches d'annotation, lexicales, syntaxiques, pragmatiques, interactionnelles et multimodales, ainsi que les difficultés à trouver un vocabulaire contrôlé précis mais sans métalangage. Ces annotations sont intégrées au logiciel libre INCEPTION (<https://inception-project.github.io/>) (Klie et al., 2018) et font l'objet d'un traitement informatique pour être visualisées facilement et de façon modulaire par des apprenants et des enseignants. Cette visualisation cherche à aider ces derniers à appréhender les interactions

dans leur dynamique et leur complexité. Les annotations peuvent également faire l'objet de requêtes pour, par exemple, visualiser toutes les demandes du corpus, toutes les salutations, toutes les ouvertures, tous les marqueurs de compréhension ou encore toutes les constructions clivées. La visualisation et l'interrogation des annotations doivent permettre de développer la compétence interactionnelle des apprenants, notamment en proposant une variété de procédés et en les exposant aux spécificités du français parlé en interaction.

Mots clés : interactions, corpus multimodaux, annotations, didactique de corpus, français langue étrangère

Références

ANDRÉ, Virginie. 2023. « De la constitution d'un corpus d'interactions à son exploitation en didactique de l'oral. Quels outils et quel accompagnement pour des pratiques innovantes ? ». *Recherches en didactique des langues et des cultures* [En ligne] 21(2). <https://doi.org/10.4000/rdlc.12651>.

ANDRÉ, Virginie. 2016. « FLEURON : Français Langue Étrangère Universitaire, Ressources et Outils Numériques. Origine, démarches et perspectives ». *Mélanges CRAPEL* 37 : 69-92.

ANDRÉ, Virginie, et Maud CIEKANSKI. 2018. « Apprendre à interagir à l'oral à partir d'un concordancier multimodal : Effets sur le développement de la conscience langagière et sur l'autonomie de l'apprenant dans le dispositif FLEURON ». C. Dejean, F. Mangenot, E. Nissen, & T. Soubrié (Éds.), *EPAL – Echanger Pour Apprendre en Ligne*. Université Grenoble Alpes.

BOULTON, Alex, & Nina VYATKINA. 2021. « Thirty years of data-driven learning: Taking stock and charting new directions over time ». *Language Learning & Technology* 25(3) : 66-89.

JOHNS, Tim. 1991. « Should you be persuaded: Two samples of data-driven learnings materials ». T. Johns & P. King (Éds.) *Classroom Concordancing*. *English Language Research Journal* Vol. 4 : 1-16.

KLIE, Jan-Christoph, Michael BUGERT, Beto BOULLOSA, Richard E. DE CASTILHO, & Iryna GUREVYCH. 2018. « The INCEPTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation ». *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*.

SURCOUF, Christian. 2022. « Le français parlé ? Eh ben j'savais pas ce que c'était ! » : Production et compréhension de la variation diaphasique en français parlé en FLE. *Mélanges CRAPEL* 1(43) : 131-156.

TRAVERSO, Véronique. 2016. *Décrire le français parlé en interaction*. Éditions Ophrys.

Notice bio-bibliographique

Virginie André est professeure en Sciences du langage à l'université de Lorraine et membre du laboratoire ATILF (UMR 7118). Elle est membre de l'équipe Didactique des langues et sociolinguistique et coordonne le groupe de recherche Langage, Travail et Formation. Elle est spécialiste de l'analyse sociolinguistique des interactions verbales et de didactique du français langue étrangère (FLE). Ses travaux portent sur l'exploitation des corpus à des fins didactiques. Elle mène ses recherches auprès des publics d'étudiants, de migrants et de natifs en insécurité langagière. Elle est responsable du projet FLEURON, un dispositif d'apprentissage qui met à la disposition des apprenants et des enseignants un corpus multimodal d'interactions authentiques interrogeable par un concordancier. Ce dispositif propose un apprentissage avec et sur corpus (<https://fleuron.atilf.fr/>).

GUIDELINES FOR THE CONSTRUCTION OF A SEGMENTED ORAL CORPUS BASED ON THE VAL.ES.CO. 3.0 CORPUS



Sara BADIA-CLIMENT

University of Valencia

sara.badia@uv.es

The construction and analysis of oral corpora represent a crucial field of research for bridging linguistics and the real language used by speakers in real contexts. Unlike written discourse corpora, which mainly present challenges related to the format of the text and the accompanying data (Torruella and Llisterra, 1999; Adolphs and Knight 2010; Carter 2013; Rojo, 2016); the oral channel, due to its immediate nature (Briz, 2010), introduces complications from the collection phase, especially in those discourses that require to be recorded secretly (Briz, 1996). These challenges are intensified when the purpose goes beyond merely representing what is said to include a detailed analysis of the discourse, such as its segmentation into discursive units (Pons 2022).

Therefore, the objective proposed by this communication is to present the processes of creation and formatting of an oral conversation corpus designed for dissemination on web platforms. This will be carried out through one of the most relevant examples of Hispanic linguistics: the Val.Es.Co. corpus (Valencian Colloquial Spanish; Pons, 2024) in its version 3.0. This corpus not only offers the transcription of its materials but also features a search system with multiple filters (characteristics of the participants and conversations, phonetics,...) that include the pragmatic units from Val.Es.Co. model. This is made possible by the tagging and segmentation work carried out on each recording of the corpus, in addition to exporting this data in a format readable by researchers.

The structure of the communication will contemplate three main segments. Initially, a brief review will be conducted of the methodology and characteristics of relevant oral corpora in the Hispanic domain, such as CORLEC (1992), APL (2017), CORPES (2018), MESA (2021), PRESEEA (2024), and AMERESCO (2024), aiming to compare their methodologies and types of discourse with those implemented in Val.Es.Co. Subsequently, the analysis will delve into the distinctive features of the conversations and the research team's objectives, linking them to the specific technical needs (Pons 2022) and the system of units adopted by Val.Es.Co. (2003; 2014). Finally, the computer solutions implemented in Val.Es.Co. 3.0 to facilitate an effective presentation of the data on the web platform will be detailed, outlining the decisions and procedures followed to achieve this objective.

In summary, this presentation aims to highlight not only the importance of oral corpora in the field of linguistics but also to provide a detailed guide for their development and analysis. It must not be

forgotten that corpus work is one of the research avenues that enable linguistics to approach the language used by speakers in real contexts (Bolaños 2015; Bernal and Hincapie 2018; García-Miguel 2022), and thanks to it, capturing and analyzing phenomena that contribute to our understanding of language is possible.

Key words: Oral corpora, Corpus construction, Val.Es.Co. 3.0, Discourse segmentation

References

- ADOLPHS, Svenja, & Ronald CARTER. 2013. Spoken corpus linguistics: From monomodal to multimodal. Routledge.
- ADOLPHS, Svenja, & Dawn KNIGHT. 2010. « Building a spoken corpus ». The Routledge handbook of corpus linguistics : 38-52.
- BERNAL-CHÁVEZ, Julio A., Moreno HINCAPIÉ, & Diana ALEJANDRA.2018. Lingüística de corpus. Bogotá: Instituto Caro y Cuervo.
- BOLAÑOS CUÉLLAR, Sergio. 2015. « La lingüística de corpus: perspectivas para la investigación lingüística contemporánea ». Forma y función 28(1) : 31-54.
- GARCÍA-MIGUEL, J. M. (2022). Lingüística de corpus. Estudios de lingüística del español, 45, 11-42.
- PONS, Salvador. 2022. Creación y análisis de corpus orales: saberes prácticos y cuestiones teóricas. Berna: Peter Lang.
- ROJO, Guillermo. 2016. « Los corpus textuales del español ». Enciclopedia lingüística hispánica 2 : 285-296.
- TORRUELLA, Joan, y Joaquim LLISTERRI. 1999. « Diseño de corpus textuales y orales ». Filología e informática. Nuevas tecnologías en los estudios filológicos : 45-77.
- VAL.ES.CO. 2003. Un sistema de unidades para el estudio del lenguaje coloquial. Oralia 6 : 7-61.
- VAL.ES.CO. 2014. Las unidades del discurso oral. La propuesta Val.Es.Co. de segmentación de la conversación (coloquial). Estudios de Lingüística del Español 35(1) : 11-71.

Biography

Sara Badia Climent is a junior research fellow involved in the DIAXX project, where she applies her expertise in corpus-building and discourse segmentation. Her research focuses on conversation analysis, digital genres, video games, and corpus construction. In 2023, she completed her doctorate with a thesis on discourse analysis in MOBA video games, demonstrating her commitment to investigating the interplay between digital discourses and natural oral conversations. As the technical director of the Val.Es.Co. 3.0 corpus, Sara has been instrumental in developing the web application and formatting the recordings to integrate both the transcription system and the segmentation model developed by the Val.Es.Co. research group.

POUR LA SEGMENTATION AUTOMATIQUE DE L'ARABE PARLÉ : L'EXEMPLE DE L'ARABE TUNISIEN



Fatma BEN BARKA-MESSAOUDI

CY Cergy Paris Université INSPE – Laboratoire EMA

fatma.messaoudi1@cyu.fr

Rayan ZIANE

Université Caen Normandie – CRISCO EA4255

rayan.ziane@unicaen.fr

Meriem BERAİK

Université Caen Normandie – CRISCO EA4255

meriem.beraik@unicaen.fr

Le traitement des données issues de l'oral est une tâche complexe, surtout quand il s'agit d'une langue privée d'orthographe usuelle. En effet, malgré le nombre croissant des travaux s'intéressant à l'étude du parler tunisien (Zribi & al., 2015 ; Ben Ahmed & al., 2018 ; Ben Barka 2018, Goncalves Teixeira & al., 2023), les tâches de transcription et d'annotation des corpus collectés sur cette langue ne cessent de soulever plusieurs interrogations, ce qui rend difficile leur exploitation linguistique.

Si plusieurs réponses ont été avancées concernant le choix du système, du mode ou bien des conventions de transcription, la question de la segmentation, indispensable à la description morphosyntaxique, n'a pas été suffisamment explorée.

Tel est le cas de la majorité des parlers arabes, l'arabe tunisien, langue fortement agglutinante, (Hamdi, 2015; Zribi 2016), se caractérise par le rattachement des clitiques aux mots simples, ce qui permet d'obtenir des formes plus complexes dites agglutinées. Séparer les clitiques de la forme fléchie est une tâche nécessaire dans une perspective de traitement des données par des outils de Traitement Automatique des Langues (TAL).

Conçues pour l'analyse de l'arabe standard moderne, les applications comme Arabic Part-of-speech Tagger (Khoja, 2001), Sakher (Chalabi, 2004) ou Aramorph (Linguistic Data Consortium, 2004) ne sont pas adaptées au traitement de nos données orales en arabe tunisien, transcrites en graphie latine sur TRANSCRIBER ou translittérées automatiquement grâce à l'API de l'outil Google Input et le translittérateur ATAR (Talafha & al., 2021). Plus récemment, on compte parmi les ressources existantes le pipeline d'annotation multi-couche (Gugliotta & al., 2020) et le corpus TARC (Gugliotta & Dinarelli, 2020), qui est propre à la modalité écrite de l'arabe parlé au Maghreb annoté en partie du discours.

Soucieux de faire avancer le débat sur les opérations d'exploration lexicale et d'étiquetage morphosyntaxique de l'arabe tunisien, nous avons choisi de segmenter un échantillon de 876 mots, soit environ 5min d'enregistrements, en 1392 morphs (Ben Barka Messaoudi & al., 2022). Ce jeu de données nous servira de base d'inférence de règles et de corpus d'apprentissage pour entraîner un système de segmentation automatique (Harrat & al., 2019) pouvant traiter notre corpus comportant 109705 mots (Ben Barka Messaoudi & al., 2023). Cette étape est déterminante et préalable aux opérations d'étiquetage grammatical et d'analyse syntaxique étant donné que les clitiques possèdent leurs propres parties du discours et fonctions syntaxiques.

En amont du traitement des unités minimales d'analyse du corpus, il convient également, pour l'analyse syntaxique en dépendance, d'effectuer une autre segmentation au niveau des unités maximales traitées. Ces séquences doivent manifester une complétude sémantique, une autonomie syntaxique et une clôture intonative. Nous montrerons, à travers l'exemple de l'arabe tunisien, comment les récentes technologies d'apprentissage machine et en particulier la diarization automatique (Plaquet & Bredin, 2023) peuvent aider au prétraitement de corpus arabes oraux.

Références

- BEN AHMED Yossra, Flora BADIN, et Linda HRIBA. 2018. « Constitution d'un corpus oral de l'arabe tunisien: une ressource essentielle pour l'étiquetage morphosyntaxique ». TALAF 2018: Traitement automatique des langues africaines (écrit et parole).
- BEN BARKA MESSAOUDI, Fatma. 2018. « Sur les enjeux méthodologiques de la construction d'un corpus d'arabe tunisien ». TALAF 2018 : Traitement automatique des langues africaines (écrit et parole). Grenoble, France. <halshs-04451827>.
- BEN BARKA MESSAOUDI, Fatma, Mustapha KHOUDRI, et Rayan ZIANE. 2022. « Le treebank comme outil de description pour les langues orales ». CEDIL22 Sciences du langage : Enjeux théoriques et pratiques méthodologiques. (hal-04095252).
- BEN BARKA MESSAOUDI, Fatma, Rayan ZIANE, et Anissa AISSANI. 2023. « L'apport des ESLO pour la documentation du continuum linguistique dans le petit Maghreb ». Actes des 11èmes Journées Internationales de la Linguistique de Corpus. (hal-04356978).
- GONCALVES TEIXEIRA, Daphne, Charle VANCAEYZEELE, et Mohamed Malek BAHRI. 2023. « Transcription Automatique de l'Arabe Parlé à Tunis : Un Pont vers l'Analyse Linguistique ». Proceedings: Actes des 5èmes journées du Groupement de Recherche CNRS « Linguistique Informatique, Formelle et de Terrain » : 135.
- GUGLIOTTA, Elisa, Marco DINARELLI, et Olivier KRAIF. 2020. Multi-task sequence prediction for Tunisian Arabizi multi-level annotation. arXiv preprint arXiv:2011.05152.
- HAMDI, Ahmed. 2015. Traitement automatique du dialecte tunisien à l'aide d'outils et de ressources de l'arabe standard: application à l'étiquetage morphosyntaxique. Thèse de doctorat. Aix-Marseille.

HARRAT, Salima, Karima MEFTOUH, et Kamel SMAÏLI. 2019. « Script Independent Morphological Segmentation for Arabic Maghrebi Dialects: An Application to Machine Translation ». *Computación y sistemas* vol. 23, no 3 :979-989.

PLAQUET, Alexis et Hervé BREDIN. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. arXiv preprint arXiv:2310.13025.

ZRIBI, Ines. 2016. Traitement automatique du dialecte tunisien: construction de ressources linguistiques. Thèse de doctorat. Université de Sfax (Tunisie).

ZRIBI, Inès, Mariem ELLOUZE, Lamia Hadrich BELGUITH, et al. 2015.« Spoken Tunisian Arabic corpus "STAC": transcription and annotation ». *Research in computing science* vol. 90 :123-135.

Notice bio-bibliographique

Fatma Ben Barka Messaoudi est enseignante-chercheuse contractuelle en sciences du langage à l'INSPÉ de l'académie de Versailles, affiliée au Laboratoire École, mutations, apprentissages (EMA) à CY Cergy Paris Université. Ses domaines de recherche comprennent l'analyse du comportement des verbes dans une langue, la linguistique de corpus, l'étude de la variation linguistique, la linguistique contrastive, les aspects généraux de la linguistique, la micro-diachronie, la documentation des langues peu dotées, l'analyse de l'arabe tunisien, la didactique du français, l'enseignement spécifique des verbes français, et la pédagogie interculturelle.

DE L'INTÉRÊT DE CORPUS DIVERSIFIÉS



Mireille BILGER

Université de Perpignan-via-Domitia – CRESEM/SLaM

bidger@univ-perp.fr

Paul CAPPEAU

Université de Poitiers – FoReLLIS

paul.cappeau@univ-poitiers.fr

Depuis les années 60, et en particulier dans le domaine franco-français, plusieurs « moments » se sont succédé dans la collecte de données exploitées en linguistique. Longtemps, un regard normatif (et traditionnel) sur la langue a conduit à privilégier les données écrites exclusivement littéraires (cf. Frantext), si ce n'est dans le domaine du FLE (cf. le Français Fondamental et plus tard le Corpus d'Orléans (ESLO1)). Une légère « modernisation » a conduit à ajouter progressivement des articles de presse « noble ». L'oral a commencé à être timidement légitimé dans les années 70. Les conditions techniques ont favorisé les enquêtes, les interviews et les productions issues des médias. Plus récemment un souci d'utiliser des productions moins sollicitées (sms pour l'écrit, interactions dans les commerces ou les repas à l'oral) a élargi les ressources disponibles. Les données disponibles de nos jours sont relativement variées tant par les caractéristiques des locuteurs que par les situations de parole ou les lieux de recueil. De plus, leur accès s'est souvent amélioré (par exemple, Gadet 2017). L'objectif de cette présentation, centrée sur les données orales, est de montrer l'intérêt du travail à partir de données diversifiées. Nous aborderons cette question à travers deux aspects :

1) La richesse des faits de langue recensés

La problématique des genres (Biber & Egbert, 2018) a mis en lumière le fait que de nombreux faits de langue se produisent plus régulièrement dans des contextes spécifiques : les passifs dans des explications, les sujets pronominaux dans des conversations, etc. Aussi se priver de certains types de production revient à regarder la langue à travers une passoire et à rater de nombreux phénomènes grammaticaux.

2) La question des unités

La question de l'unité d'analyse se pose différemment selon les cas. Ainsi, dans des conversations par exemple, les échanges sont souvent brefs et les énoncés peuvent, sur le plan syntaxique, sembler elliptiques ou rarement complets. La notion de locuteur collectif (Loufrani, 1984) permet de dépasser le cadre du tour de parole et de mieux saisir l'organisation des échanges. Dans

d'autres productions comme les explications ou les passages argumentatifs qui reposent sur des procédés syntaxiques de reprise et d'amplification, c'est plutôt une séquence plus développée que la « phrase » qui constitue l'unité la plus appropriée (cf., entre autres, Benveniste 2010, Debaisieux 2013).

Le lien entre diversité des données et variété des unités d'analyse pertinentes, moins traité, sera au cœur de notre présentation et sera développé à travers des exemples empruntés à plusieurs situations de parole.

Références

- BENVENISTE, Claire. 2010. *Le français : Usages de la langue parlée*. Leuven. Peeters.
- BIBER, Douglas & Jesse EGBERT. 2018. *Register Variation Online*. Cambridge. Cambridge University Press.
- CHENG, Winnie. 2012. *Exploring Corpus Linguistics. Language in Action*. Abingdon. Routledge.
- DEBAISIEUX, Jeanne-Marie, 2013. *Analyses linguistiques sur corpus, subordination et insubordination en français*. Paris. Hermès-Lavoisier.
- GADET, Françoise (éd). 2017, *Les parlers jeunes dans l'Île-de-France multiculturelle*. Ophrys. Paris.
- GOUGENHEIM, Georges, Paul RIVENC & Aurélien SAUVAGEOT. 1956. *L'élaboration du français élémentaire : étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Didier. Paris.
- LOUFRANI, Claude. 1984. Le locuteur collectif. Typologie de configurations discursives. *RSFP*. 6 : 169-193.

Notice bio-bibliographique

Mireille Bilger, professeur de sciences du langage, membre du laboratoire CRESEM/ SLaM (Sociétés, Langages, Mobilités) de l'université de Perpignan. Ses recherches portent sur la syntaxe du français parlé et écrit. La constitution, l'édition et l'exploitation des corpus oraux sont également des thèmes développés dans plusieurs de ses travaux.

Paul Cappeau, professeur de sciences du langage associé au laboratoire FoReLLIS (Poitiers). Ses publications et ouvrages portent principalement sur la constitution et l'exploitation de corpus oraux ainsi que sur l'étude des productions écrites d'enfants de l'école primaire. Il est l'auteur de *Une grammaire à l'aune de l'oral* (2021, PUR).

LA PAROLE DES MÉDECINS : UN ORAL SOUS L'INFLUENCE DE L'ÉCRIT ?



Mylène BLASCO

Université Clermont-Auvergne – LRL UR999

mylene.blasco-dulbecco@uca.fr

Paul CAPPEAU

Université de Poitiers – FoReLLIS

paul.cappeau@univ-poitiers.fr

Notre communication porte sur la linguistique de l'oral à travers l'analyse d'un corpus d'interactions spécialisées : des consultations médicales. Il s'agit de vérifier dans quelle mesure, dans ces discours spécifiques, la syntaxe de l'oral peut être affectée par un environnement écrit prégnant.

Il n'est pas rare de trouver des « traces » de l'écrit dans la parole, on peut penser à certaines expressions comme « entre guillemets » ou « point final » qui renvoient à l'usage graphique (Blanche-Benveniste, 1997). Dans certains usages de langue professionnelle (hommes politiques, enseignants...) l'oral prend souvent appui sur des documents écrits qui peuvent avoir une incidence sur les formulations orales. On propose ici d'analyser, à partir d'un corpus de consultations authentiques recueillies à l'hôpital, la parole de médecins et notamment d'observer comment l'écrit - aux niveaux lexical et syntaxique - s'immisce dans l'échange.

Notre travail prendra appui sur l'usage des noms déverbaux (1), certaines caractéristiques de la syntaxe verbale (2) et la complexité du système nominal (3) :

(1) une altération de l'articulation (CO-A2)

(2) ça vous permettra d'aider à la gestion de l'alimentation (CO-B1)

(3) des mutations + effectivement sur un gène + qu'on appelle le gène H.F.E qui était responsable de l'hémochromatose (CO11)

Cette présence de formes propres à l'écrit modèlerait les habitudes de pensée des médecins pendant leurs études, partant du principe que la maîtrise d'une langue à caractère médical est une compétence primordiale pour supporter le raisonnement clinique et asseoir un statut d'expert lors de la consultation. Les médecins auraient plus ou moins de mal à s'écarter de ces patterns, au risque de nuire à l'efficacité de la communication et de provoquer, parfois, des difficultés de compréhension chez le patient.

Mots clés : linguistique de l'oral, corpus, syntaxe, interactions médecins-patients.

Références

- BLANCHE-BENVENISTE, Claire. 1997. *Approches de la langue parlée en français*. Paris. Ophrys.
- BLASCO, Mylène. (Éd.) 2024 (sous presse). *Langage, langue, parole dans la relation de soin*, PUBP : Clermont-Ferrand.
- COSNIER Jacques. 1993. « Les interactions en milieu soignant ». J. Cosnier, M. Grosjean, & M. Lacoste (éds.) *Soins et communication : approches interactionnistes des relations de soins* : 17-32. Lyon : PUL.
- HALLIDAY, Michael A.K. 1985. *Spoken and written Language*. Oxford. Oxford University Press.
- VECCHIATO, Sara, et Sonia V. GEROLIMICH. 2013. « La langue médicale est-elle 'trop complexe' ? ». *Nouvelles perspectives en sciences sociales* 9(1) : 81-122. <https://doi.org/10.7202/1024039ar>.
- ROUGET, Christine. 2000. « Les nominalisations sont-elles réservées aux descriptions techniques ? ». Bilger M. (éd), *Corpus – Méthodologie et applications linguistiques*. Paris, Honoré Champion : 296-305.
- VERGELY, Pascale, Anne CONDAMINES, Cécile FABRE, Amélie JOSSELIN-LERAY, Josette REBEYROLLE, et Ludovic TANGUY. 2009. « Analyse linguistique des interactions patient/médecin ». Felix, C. & Tardif, J. (dir.). *Actes éducatifs et de soins*, Nice (publication numérique : <halshs- 00532967>).

Notice bio-bibliographique

Mylène Blasco est professeure des universités en sciences du langage à l'Université Clermont-Auvergne et sa recherche au Laboratoire de Recherche sur le Langage (LRL). Ses travaux portent sur la syntaxe du français parlé dans les interactions notamment professionnelles. Elle a assuré la responsabilité scientifique du projet structurant DECLICS (Dispositifs d'études cliniques sur les corpus santé, Région AURA de 2016 à 2021) et est responsable, depuis 2019, du DIU Relations médecins-soignants/patients de l'UCA. Ses derniers ouvrages sont *Parler à l'hôpital* (éd.) (2022), *Nodus Publicationem : Munster et Langage, langue, parole dans la relation de soin* (éd) (sous presse 2024), PUBP : Clermont-Ferrand.

Paul Cappeau est professeur des universités en sciences du langage associé au laboratoire FoReLLIS (Poitiers). Il a fait paraître de nombreuses publications et ouvrages dévolus à la constitution et à l'exploitation de corpus oraux et à l'étude des productions écrites d'enfants de l'école primaire. Il est l'auteur de *Une grammaire à l'aune de l'oral* (2021, PUR). Il a contribué au projet DECLICS et a participé à l'écriture des ouvrages *Parler à l'hôpital* M. Blasco (éd.), 2022, *Nodus Publicationem : Munster et Langage, langue, parole dans la relation de soin*, M. Blasco (éd.), 2024 (sous presse), PUBP : Clermont-Ferrand.

L'ANALYSE SYNTAXIQUE EN « GRILLES » DES TEXTES ORAUX : ILLUSTRATIONS AVEC DES DONNÉES DE FINNOIS ET D'HÉBREU MODERNE



Outi DUVALON

Inalco – SeDyL

outi.duvallon@inalco.fr

Il-Il YATZIV-MALIBERT

Inalco – CERMOM

il-il.yatziv-malibert@inalco.fr

Cette communication se propose de se pencher sur la question de la structuration de l'oral spontané en termes syntaxiques. Il s'agira de partir du cadre théorique et méthodologique élaboré depuis la fin des années 1970 par l'équipe du Groupe aixois de recherche en syntaxe (GARS) pour l'étude grammaticale du français parlé (par ex. Blanche-Benveniste et al. 1979, 1990) et d'interroger la pertinence des outils d'analyse mis en place à la lumière de leur application aux données de langues éloignées du français, comme le finnois et l'hébreu moderne.

Seront présentés d'abord les choix méthodologiques et les fondements théoriques de l'analyse syntaxique en « grilles » (Blanche-Benveniste 1990), notamment le refus de procéder à un découpage des textes bruts en segments phrastiques et la distinction de deux types de regroupements d'unités dans la chaîne parlée linéaire, à savoir les suites syntagmatiques et les listes paradigmatisées. Nous examinerons ensuite des données tirées de corpus oraux spontanés de finnois et d'hébreu moderne dans le but de mettre l'accent sur une vision dynamique de l'exploitation de l'axe paradigmatique correspondant à des emplacements syntaxiques organisés par un élément recteur. L'attention sera portée aussi bien sur des aspects grammaticaux, comme par exemple les indices formels permettant d'identifier les listes paradigmatisées à distance, que des aspects interactionnels, tels que la co-construction des listes paradigmatisées. Enfin, en nous appuyant sur des exemples d'autocorrections, de reformulations et de reprises d'énoncés interrompus, nous discuterons des perspectives qu'ouvre cette approche donnant à voir une syntaxe « on-line » (Auer 2009) sur le savoir grammatical des locuteurs tel qu'il sous-tend leurs productions spontanées.

Mots clés : analyse syntaxique en «grilles, syntaxe on-line, liste paradigmatique, co- construction, reformulations, finnois, hébreu moderne

Références

AUER, Peter. 2009. « On-line syntax: Thoughts on the temporality of spoken language ». *Language Sciences* 31 : 1-13.

BLANCHE-BENVENISTE, Claire. 1990. « Un modèle d'analyse syntaxique "en grilles" pour les productions orales ». *Anuario de Psicologia* 47 : 11-28.

BLANCHE-BENVENISTE, Claire, Bernard BOREL, José DEULOFEU, Jacques DURAND, Alain GIACOMI, Claude LOUFRANI, Boudjema MEZIANE, et PAZERY Nelly. 1979. « Des grilles pour le français parlé ». *Recherches sur le français parlé* 2 :163-205.

BLANCHE-BENVENISTE Claire, Mireille BILGER, Christine ROUGET, et Karel VAN DEN EYNDE. 1990. *Le français parlé. Études grammaticales*. Paris : Éditions du CNRS.

Notice bio-bibliographique

Outi Duvallon est professeur de linguistique finnoise à l'Inalco et membre de l'UMR SeDyL.

Il-Il Yatziv-Malibert est professeur de linguistique hébraïque à l'Inalco et membre de l'équipe CERMOM.

PARAMÉTRAGE ET ANALYSE D'UN CORPUS ORAL MODÉLISÉ EN LV2 RUSSE : ÉTUDE À DES FINS PÉDAGOGIQUES

Kirill GANZHA

Université Paris Nanterre – CRPM

kirill.ganzha@parisnanterre.fr

En observant les premières productions orales, tant spontanées qu'induites, des grands débutants francophones qui s'expriment en russe en classe, nous avons remarqué un certain nombre d'erreurs d'intonation récurrentes. Par exemple, nous avons proposé de lire une conversation en famille : le père rentre du travail, voit sur la table une coupe de fruits et demande à sa fille si c'est elle qui a acheté les pommes :

- Doč', ty kupila âbloki?
- Net, Marina kupila.
- Litt. : - Fille, tu as-acheté pommes ?
- Non, Marina as-acheté.
- Trad. : - Ma chérie, c'est toi qui as acheté les pommes ?
- Non, c'est Marina (qui les a achetées).

En écoutant les productions, nous constatons que dans la première réplique certains étudiants se trompent dans le choix du terme focalisé (toi vs Marina) et accentuent le verbe (kupila « as acheté ») avec un éventuel écho accentuel sur l'ultime syllabe droite de l'objet (*âbloki), ce qui prive le dialogue de tout sens. D'autres arrivent à identifier sur quel mot porte l'accent nucléaire de chaque réplique, mais ne réussissent pas à le réaliser correctement, sur la syllabe porteuse de l'accent lexical, comme dans *M:arina, où, marqué par un allongement consonantique, l'accent remonte indûment sur l'extrémité gauche du terme focalisé.

Les apprenants se heurtent ainsi principalement à deux problèmes : l'identification de la trajectoire de la ligne mélodique, conditionnée par la modalité de l'énoncé et la distribution des accents (aspect cognitif), et la réalisation d'une courbe mélodique mentalement préconstruite (aspect cognitif et articulatoire). Le deuxième problème, pour sa part, tient à la question de l'assimilation du rythme des mots, celui-ci étant en grande partie conditionné par l'accent lexical (hiérarchisation des niveaux prosodiques). Ces faits nous ont permis de cerner les éléments prosodiques qui nous semblent les plus importants et les plus difficiles à acquérir en russe pour un francophone, de préciser notre approche analytique et de paramétrer notre corpus en termes de segmentation et d'annotation.

Dans une visée didactique – entraînement aux schémas mélodiques et accentuels de l'oral spontané chez les étudiants – nous avons modélisé des situations monologiques et dialogales non spécialisées (small talk). Nous avons proposé à 12 francophones de lire à voix haute 4 conversations et nous avons récupéré un corpus son de 50 minutes environ.

Lors de l'élaboration et de l'analyse du corpus, nous avons été amené à réfléchir et à trancher sur un certain nombre de questions, parmi lesquelles :

- o le choix entre parole spontanée et parole lue en LV2 : problème du manque de naturel ;
- o la recherche de l'équilibre entre les conditions techniques et un comportement corporel naturel : problème débouchant sur la fiabilité de certains paramètres acoustiques, la hauteur par exemple ;
- o la définition des niveaux de segmentation, pouvant aller jusqu'aux évènements acoustiques, et leur annotation : choix entre informativité et économie ;
- o le choix des paramètres acoustiques pertinents : soumission du corpus à l'audition d'un groupe de locuteurs natifs et traitement de leurs réponses.

Références

BAQUÉ, Lorraine. 1998. « La prosodie des productions d'apprenants de russe langue étrangère ». Comtet, Roger (éd.), Billières, Michel (éd.). Autour du russe : études perceptives et comparatives. Slavica occitania. Toulouse 6 : 27-53.

BAUDE, Olivier, Claire BLANCHE-BENVENISTE, Marie-France CALAS, et al. 2006. Corpus oraux, guide des bonnes pratiques. CNRS Éditions. Presses Universitaires Orléans.

BILLIÈRES, Michel. 1985. Contribution à l'étude des erreurs de prononciation des russophones apprenant le français en vue de l'application à des pratiques de phonétique corrective. Thèse pour le Doctorat du 3e cycle en Science du Langage. Université de Toulouse-Le Mirail (Toulouse II).

BORRELL André, et Jeanne SALSIGNAC. 2002. Chapitre 4 : Importance de la prosodie en didactique des langues (application au FLE). Renard, Raymond (éd.). Apprentissage d'une langue étrangère/seconde. Vol. 2 La phonétique verbo-tonale. 1e édition. Pédagogies en développement. De Boeck Supérieur, Bruxelles, Belgique :163-182.

DELAIS-ROUSSARIE, Elisabeth. 2002. « Constituer des corpus oraux : Méthodes et Outils ». Carnets de grammaire. Rapports internes de l'ERSS. Équipe de Recherche en Syntaxe et Sémantique, UMR 5610. CNRS & Université de Toulouse-Le Mirail. Toulouse, France.

DELAIS-ROUSSARIE, Elisabeth, et Brechtje POST. 2014. Corpus annotation and transcription systems. Durand, Jacques (ed.). The Oxford Handbook of Corpus Phonology. Oxford University Press :1-21.


FOUGERON, Irina. 1986 a. Analyse prosodique de la phrase russe : l'organisation du message dans les phrases assertives à un, deux ou trois composants. Thèse pour le Doctorat d'État, Université de la Sorbonne Nouvelle Paris III.

Mots clés : accent lexical, accent nucléaire, rythme du mot, prosodie, intonation, Russe Langue Étrangère, erreurs des francophones en russe, transfert, didactique

Notice bio-bibliographique

Ganzha Kirill, maître de conférences en langue russe au Département des Études slaves à l'Université Paris Nanterre ; membre titulaire du CRPM (Centre de recherches pluridisciplinaires multilingues) et membre associé du laboratoire SeDyL (Structure et Dynamique des Langues). Sa thèse de doctorat en Sciences du langage, intitulée « Production des accents lexical et nucléaire russes par les francophones : étude expérimentale descriptive et analytique » a été réalisée sous la direction de Christine Bonnot et Claire Pillot-Loiseau et soutenue en 2023. Il poursuit ses recherches sur l'aspect prosodique du russe contemporain et s'intéresse en outre aux variations de l'ordre des mots.

SEGMENTATION DU DISCOURS DE DÉBAT EN ANGLAIS LANGUE ÉTRANGÈRE. COMMENT IDENTIFIER L'UNITÉ SPONTANÉE DANS L'ORAL PRÉPARÉ ?



Marina HAAN

Toulouse 3 Paul Sabatier – LAIRDIL

marina.haan@iut-tlse3.fr

Dans cette communication, je propose une réflexion sur l'identité de l'unité spontanée dans le discours préparé de débat en ALVE (secteur LANSAD). En effet, comment identifier (reconnaître et annoter) l'unité spontanée dans le discours qui a été produit (écrit par l'apprenant en amont) et qui a été « réactualisé » ou « ré-évoqué » (Foucault, 1969 : 149) en oral, lors de la prise de parole en public ? Autrement dit, comment peut-on segmenter le discours monologal, quand il s'agit de la « coexistence », dans le système épistémique du locuteur, de 2 textes similaires : le texte planifié (le support du discours préparé avant la prise de paroles en public) et le texte prononcé, qui a été spontanément modifié (adapté, complètement ou partiellement, par ce locuteur à la situation réelle de communication) ?

L'analyse de nombreux discours de débat en cours LANSAD, faisant partie de l'enquête de terrain au sein du département GEA - IUT Toulouse 3, m'a permis de repérer au moins 4 propriétés identificatrices de l'unité discursive spontanée que je souhaiterai présenter lors de ce colloque international :

- le niveau d'improvisation (mots, syntagmes, phrases, énoncés) ;
- l'opération linguistique d'improvisation (l'opération à travers laquelle l'unité spontanée émerge dans le discours préparé) ;
- le cadre prosodique (notamment, les pauses qui bornent l'unité du discours) ;
- le cadre gestuel (plus particulièrement, l'ampleur et le rythme des gestes co-verbaux).

Le repérage de ces propriétés s'inscrit dans la théorie des systèmes dynamiques (Larsen-Freeman et al., 2007 ; Dörnyei et al., 2015 ; Toffoli, 2020) et la théorie de l'enaction (Varela et al., 1993 ; Aden, 2017 ; Bottineau et al., 2017) adaptées au contexte de l'apprentissage des langues (perspective didactique).

Dans le cadre de cette communication, je tiens également à expliquer et à démontrer empiriquement pourquoi il est important de prendre en compte la gestualité du locuteur dans la segmentation de son discours. Ma réflexion s'appuie sur l'idée de Larsen-Freeman et al. (2007) que le discours représente un système dynamique complexe. Régie par son créateur (l'orateur), ce système fonctionne sur deux registres langagiers (verbal et co-verbal) qui, à leur tour, s'organisent en deux sous-systèmes en interaction. On ne peut donc pas segmenter le discours oral (différencier les unités spontanées des unités non spontanées) sans prendre en compte les relations entre les

composantes linguistiques (les éléments verbaux) et les composantes extralinguistiques (les éléments non verbaux) co-émergeant, à un moment donné, dans le système du discours oral.

Par conséquent, on peut supposer qu'il existe certaines corrélations (régularités de co-émergence) entre les composantes linguistiques et extralinguistiques (plus particulièrement, les conduites gestuelles du locuteur) associées aux unités spontanées du discours (ceci par rapport aux unités non spontanées). Les résultats de l'enquête de terrain prouvent que les corrélations évoquées se manifestent, dans les débats en anglais, très fréquemment, plus particulièrement au niveau de l'ampleur et du rythme gestuel des orateurs. Je constate que la connaissance de ces corrélations contribue à la distinction entre les unités spontanées (improvisées) et les unités préparées (reproduites) dans le discours oral, ce qui peut faciliter en général la segmentation de ce dernier (surtout en l'absence des supports écrits pour l'analyse comparative des contenus). Ainsi, l'implication du corps dans le comportement langagier de l'orateur (l'improvisation vs la reproduction du texte préparé) valide l'inclusion du cadre gestuel dans l'ensemble évoqué des propriétés qui permet d'identifier l'unité spontanée dans le discours oral préparé.

Mots clés : discours de débat, unité spontanée du discours, système dynamique, enaction, LANSAD

Références

ADEN, Joëlle. 2017. « Langues et langage dans un paradigme éactif ». *Recherches en didactique des langues et des cultures*, n° 14, vol. 1. [<http://journals.openedition.org/rdlc/1>], consulté le 27 mai 2021.

BOTTINEAU, Didier, et Michaël GRÉGOIRE. 2017. « Le langage humain, les langues et la parole du point de vue du languaging et de l'éaction ». *Intellectica (Langage et éaction : corporéité, environnements, expériences, apprentissages)* n° 68 :7-15.

DÖRNYEI, Zoltan, Peter D. MACINTYRE, et Alastair HENRY. 2015. *Motivational dynamics in language Learning*. Bristol : Multilingual Matters.

FOUCAULT, Michel. 1969. *L'archéologie du savoir*. Paris :Gallimard.

LARSEN-FREEMAN, Diane, et Lynne CAMERON. 2007. *Complex systems and applied linguistics*. Oxford : Oxford University Press.


TOFFOLI, Denyze. 2020. *Informal learning and institution-wide language provision: University language learners in the 21st Century*. Palgrave Macmillan.

VARELA, Francisco, Evan THOMSON, et Eleanor ROSCH. 1993. *L'inscription corporelle de l'esprit*. Paris : Le Seuil.

Notice bio-bibliographique

Marina Haan est maître de conférences en didactique de l'anglais au département de la Gestion des Entreprises et des Administrations (GEA) à l'IUT Ponsan, Université Toulouse 3 Paul Sabatier, et membre titulaire du laboratoire inter-universitaire de recherche en didactique des langues (LAIRDIL) depuis 2016. Elle est auteur de l'ouvrage *Systèmes dynamiques et émergence du discours spontané en anglais. Apprendre à improviser dans un débat en langue étrangère* (Presses universitaires de Rennes) qui paraîtra début 2025.

DONC EUH VOILÀ DONC JE MANQUE DE VOCABULAIRE. PAUSES REMPLIES ET RÉPÉTITIONS EN TANT QUE STRATÉGIES EMPLOYÉES POUR ORGANISER LE DISCOURS SPONTANÉ À L'EXEMPLE DU MINI-CORPUS DE FRANÇAIS PARLÉ EN POLOGNE



Karolina KOPCZYŃSKA

Université Adam Mickiewicz de Poznań

karolina.b.kopczynska@gmail.com

L'objectif de la présente communication consiste à présenter la méthodologie qui a permis d'établir notre propre corpus du français parlé, mais également à vérifier les tendances prosodiques et combinatoires des disfluences à l'exemple des pauses remplies et des répétitions. Les données collectées contiennent 150 minutes d'enregistrements audio, ce qui correspond à la transcription de 28354 mots. Le matériel linguistique a été recueilli à l'aide des trois entretiens semi-guidés, avec deux personnes (des locutrices et des locuteurs natifs, francophones, habitant en Pologne) et transcrit à la main selon les conventions proposées par le groupe ICOR (Canut et al., 2018).

Nous avons analysé et annoté 737 occurrences de *eah*. Cet élément langagier, faisant partie intégrante du français oral spontané, est décrit dans la littérature comme une forme d'hésitation, une pause remplie ou encore une auto-interruption (Candea, 2000 ; Duez, 2001 ; Pallaud, Bertrand, 2020). Au cours des analyses, nous avons focalisé notre attention sur la position de *eah* dans les énoncés, ses aptitudes combinatoires, ses fonctions et ses propriétés prosodiques, à savoir le débit et la fréquence fondamentale. Cette dernière a été calculée à l'aide du logiciel Praat (Boersma, Van Heuven, 2001). Nous avons constaté que *eah* se combine avec d'autres marqueurs discursifs, tels que par exemple : *donc*, *ben*, *enfin*, ce qui a déjà fait l'objet de diverses recherches (Candea, 2000 ; Duez, 2001). Parmi les occurrences collectées, nous avons observé que cette manifestation d'hésitation a tendance à se reproduire plusieurs fois au niveau d'un même énoncé. À l'exemple de notre corpus, nous avons également remarqué qu'après la pause remplie, les interlocuteurs répètent une partie de la syntagme déjà prononcée ou abandonnent le contenu qu'ils ont commencé en faveur d'une nouvelle idée (Pallaud, Bertrand, 2020). L'utilisation des pauses remplies est une stratégie qui permet à un locuteur de réfléchir son énoncé en garantissant du temps pour une recherche de la syntagme adéquate ou pour l'encodage du contenu présenté par l'interlocuteur (Duez, 2001 ; Dodane, Hirsch, 2018).

Mots clés : corpus oral, français parlé, disfluence, pause remplie, répétition

Références

- BAUDE, Olivier, et al. 2006. Corpus oraux, guide des bonnes pratiques 2006. Orléans : CNRS Editions, Presses Universitaires Orléans.
- BLANCHE-BENVENISTE, Claire. 2010. Approches de la langue parlée. Paris : Ophrys.
- BOERSMA, Paul, et Vincent VAN HEUVEN. 2001. « Speak and unSpeak with PRAAT ». *Glott International* 5(9/10) : 341-347.
- CANDEA, Maria. 2000. Contribution à l'étude des pauses silencieuses et des phénomènes dits « d'hésitation » en français oral spontané. Thèse de doctorat. Université de Paris III.
- CANUT, Cécile, et al. 2018. Le langage, une pratique sociale : Éléments d'une sociolinguistique politique. Besançon: Presses universitaires de Franche-Comté.
- DODANE, Christelle, et Fabrice HIRSCH. 2018. « L'organisation spatiale et temporelle de la pause en parole et en discours ». *Langages* 211 : 5-12.
- DUEZ, Danielle. 2001. « Signification des hésitations dans la production et la perception de la parole spontanée ». *Parole* 17 : 113-137.
- DUEZ, Danielle. 2001. « Caractéristiques acoustiques et phonétiques des pauses remplies dans la conversation en français ». *Travaux Interdisciplinaires du Laboratoire Parole et Langage* 20 : 31-48.
- HENRY, Sandrine, Estelle CAMPIONE, et Jean VÉRONIS. 2004. « Répétitions et pauses (silencieuses et remplies) en français spontané ». *Actes des XXV^e Journées d'Etude sur la Parole (JEP'2004)* : 1-4.
- PALLAUD, Berthille, et Roxanne BERTRAND. 2020. « Auto-interruptions et disfluences à l'oral. C'était euh tu vois complètement loufoque comme si ouais euh comme situation ». *Manuel de pausologie. Recueil de recherches sur la parole et le discours, Langue et Parole* : 21-47.

Notice bio-bibliographique

Karolina Kopczyńska est titulaire des diplômes de master en psychologie et en philologie romane (Université Adam Mickiewicz de Poznań). Lauréate du concours Study@research réalisé à l'Université Adam Mickiewicz de Poznań avec le projet visant la création du corpus de français parlé en Pologne. Stagiaire au laboratoire de recherche THÉMA (au sein de la COMUE Bourgogne Franche-Comté) en 2023 où elle a assisté au projet intitulé *De quoi parle la presse française lorsqu'elle dit « exode urbain » ? D'une revalorisation symbolique des campagnes jusqu'aux discours accompagnant leur réintroduction dans le marché : une transition discursive majeure*. Ses recherches portent sur la linguistique de corpus, la lexicométrie, la textométrie et le bilinguisme.

MÉTHODOLOGIE(S) DE SEGMENTATION DES TRANSCRIPTIONS DE L'ORAL EN VUE DE LA CRÉATION D'UNE MESURE D'ÉVALUATION DES PRODUCTIONS VERBALES ENFANTINES

Loïc LIÉGEOIS

Université Paris Cité – LLF UMR 7110,
CLILLAC-ARP EA 3967
loic.liegeois@u-paris.fr

Christine DA SILVA-GENEST

Université de Lorraine – Modyco UMR 7114,
DevAH UR 3450
christine.da-silva-genest@univ-lorraine.fr

Christophe BENZITOUN

Université de Lorraine – ATILF UMR 7118
christophe.benzitoun@atilf.fr

Caroline MASSON

Université Paris Sorbonne Nouvelle
CLESTHIA EA 7345
caroline.masson@sorbonne-nouvelle.fr

Christophe PARISSÉ

Université Paris Nanterre – INSERM,
Modyco UMR 7114
cparisse@parisnanterre.fr

L'évaluation des compétences langagières des enfants se révèle être un processus complexe faisant appel à une grande diversité de méthodes oscillant entre approches formelles et écologiques. L'objectif principal de notre projet est de fournir des outils fiables et objectifs pour évaluer des productions verbales orales d'enfants à développement typique ou atypique âgés entre 5 et 7 ans en situations naturelles (anonyme, 2023). L'un des axes de travail consiste en la création d'un outil de mesures automatiques de la complexité des productions enfantines, réalisées à partir de transcriptions d'interactions naturelles parent-enfant en situation de jeu libre.

La définition de la complexité linguistique et le choix des caractéristiques linguistiques à observer ont été des tâches ardues auxquelles nous avons été confrontés. Ainsi, nous avons construit une typologie du degré d'élaboration du discours fondée sur des caractéristiques morphologiques (temps verbaux), syntaxiques et lexicales. Cette typologie a été élaborée à partir d'une analyse manuelle et qualitative de productions enfantines transcrites. Nous avons ensuite produit une chaîne de traitements permettant d'obtenir des scores de mesure de complexité à partir de cette typologie :

- l'analyse syntaxique des productions enfantines au moyen de l'outil Stanza (Qi et al., 2020) ;
- le relevé automatique, dans les fichiers en sortie de l'analyse syntaxique, des différents indices de complexité.

Très tôt, l'impact de la segmentation des productions enfantines était patente. Pour un texte écrit, l'analyse automatique de la complexité peut facilement prendre appui sur la segmentation en phrases (voir par exemple Battistelli et al., 2022). S'agissant de transcriptions de productions orales, les mesures se trouvent particulièrement sensibles à la méthode de segmentation choisie. Les méthodes de segmentation des productions orales varient fortement en fonction des méthodologies de transcription mises en place et des outils utilisés. Nous distinguons deux méthodes principales : la segmentation en tours de parole et celle en énoncés. Cette dernière est la plus souvent employée pour l'étude des productions enfantines (Parisse et Lenormand, 2006) car elle permet notamment d'obtenir une mesure de complexité simple mais largement utilisée dans le domaine de l'acquisition du langage, à savoir la longueur moyenne des énoncés (ou LME, MLU en anglais, Brown, 1973).

Nous proposons dans cette communication de comparer les résultats obtenus pour la mesure de complexité des productions enfantines en fonction des méthodes de segmentation mises en oeuvre. Notre communication s'articulera autour de trois parties.

Dans un premier temps, nous présenterons la notion de mesure de complexité des productions enfantines en mettant l'accent sur les caractéristiques linguistiques retenues afin de l'évaluer.

La partie principale sera ensuite consacrée à la comparaison des mesures obtenues en fonction de la segmentation paramétrée pour l'analyse morphosyntaxique via Stanza : segmentation manuelle en énoncés et segmentation automatique du flux continu de la parole opérée par l'outil.

Enfin, nous insisterons sur l'intérêt que pourrait avoir à terme un tel outil d'évaluation automatique de la complexité du discours des enfants pour les orthophonistes.

Références

BATTISTELLI, Delphine, et al. 2022. « Une chaîne de traitements pour prédire et appréhender la complexité des textes pour enfants d'un point de vue linguistique et psycho-linguistique ». Traitement Automatique des Langues Naturelles (TALN 2022), Jun 2022. Avignon, France :236-246. [hal-03701501].

BROWN, Roger. 1973. « Development of the first language in the human species. » *American Psychologist* 28(2): 97-106. [<https://doi.org/10.1037/h0034209>].

DA SILVA et al., à paraître. Évaluation du langage oral: perspectives pour la recherche et la clinique. *Études de linguistique appliquée : revue de didactologie des langues-cultures*, 210 : 33-140. [hal-04250436].

PARISSE, Christophe, et Marie-Thérèse LE NORMAND. 2007. « Une méthode pour évaluer la production du langage spontané chez l'enfant de 2 à 4 ans ». *Glossa* 97 : 10-30. [halshs-00165632].

QI, Peng, et al. 2020. « Stanza: A Python natural language processing toolkit for many human languages ». Savova et al. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Notice bio-bibliographique

Loïc Liégeois est ingénieur de recherche à l'Université Paris Cité. Il est spécialisé en recueil, structuration et analyse de corpus linguistiques et en acquisition du langage.

Christine da Silva-Genest est maîtresse de conférences en sciences du langage à l'université de Lorraine. Elle bénéficie actuellement d'une délégation CNRS au laboratoire Modyco (UMR 7114). Elle est spécialisée en acquisition du langage. Ses domaines de recherche concernent notamment le développement typique et atypique du langage. Son domaine d'expertise s'étend également aux méthodes de remédiation linguistique.

Christophe Benzitoun est maître de conférence en sciences du langage à l'université de Lorraine. Il est notamment responsable de l'équipe De la Syntaxe au Discours au sein du laboratoire ATILF (UMR 7118). Ses domaines de recherche concernent notamment la syntaxe de l'oral et la linguistique de/sur corpus.

Caroline Masson est maîtresse de conférences à l'Université Sorbonne Nouvelle. Ses domaines de recherche concernent notamment l'acquisition du langage et les modalités d'interaction adulte/enfant à développements typique et atypique. Elle possède également une expertise en recueil et analyse de corpus oraux recueillis en situations clinique et éducative.

Christophe Parisse est chargé de recherches à l'INSERM. Ses domaines de recherche concernent notamment l'acquisition du langage et la linguistique de/sur corpus. Il est par ailleurs directeur du laboratoire Modyco (UMR 7114) et coordinateur du consortium CORLI (Huma-Num).

INTERACTIONS DANS LE CONTEXTE DU POLYHANDICAP



Karine MARTEL

INSEI – Grhapes

karine.martel@inshea.fr

Le polyhandicap se caractérise par plusieurs besoins complexes qui influent sur le quotidien des personnes. Parmi ces besoins, les difficultés à communiquer sont particulièrement marquées. La majorité des personnes polyhandicapées ont du mal à se faire comprendre puisqu'elles se retrouvent quasiment toutes sans langage oral signifiant ou sans outils de CAA. Par là-même, elles ont du mal à comprendre les messages qui leur sont adressés. Elles déploient des stratégies singulières dont l'identification et l'accès demeurent subtils. En outre, les compétences cognitives et corporelles limitées des personnes polyhandicapées, le degré de connaissance et leur perception du ou des partenaires, le contexte des échanges, ainsi que leur personnalité impactent les composantes au cœur du fonctionnement des interactions, à savoir la sensibilité, l'attention conjointe, la co-régulation et les émotions (Hostyn et Maes, 2009). Le but de l'étude est de s'intéresser aux formes discursives utilisées par les accompagnants pour tenter de repérer s'il existe des mécanismes sous-jacents à la dynamique dialogique entre ces partenaires aux compétences très différentes.

Le travail porte sur les répétitions formulées par les professionnels en charge des séances classe dans les institutions accueillant des jeunes polyhandicapés. Les répétitions ont été très étudiées dans le domaine de l'acquisition du langage, notamment des premiers échanges adultes/jeunes enfants. Pour certains auteurs, il s'agit de la « première relation de type dialogique, le premier lien de continuité intradiscursive [...] » (p. 31, Bernicot, Salazar-Orvig et Veneziano, 2006).

Le présent travail a été réalisé à partir du corpus du projet Polycom dont l'objectif était de mieux comprendre les processus de communication au sein du groupe classe, entre pairs, et entre élève avec polyhandicap et adulte à partir d'une analyse des vidéos recueillies dans le projet Polyscol (Toubert-Duffort et al., 2018). Deux jeunes de la cohorte (comprenant dix participants au total) ont été retenus, respectivement âgés de 14 ; 1 ans et 10 ; 2 ans, et ayant reçu un diagnostic de polyhandicap fixé congénital. Ces derniers sont capables d'établir un contact visuel et de communiquer au moyen de mimiques et de vocalises mais, ne produisent pas de langage verbal.

Un inventaire d'auto-répétitions, inspiré de l'étude de Bocéréan et collaborateurs (2010), a été constitué pour coder les prises de parole des enseignants sachant que les hétéro-répétitions n'ont pas lieu d'être. Les données révèlent que ces adultes produisent différents types d'auto-répétitions (avec ajout, partielles, etc.) mais, privilégient les auto-répétitions identiques qu'ils réalisent avec une durée réduite de 30% par rapport à la réplique initiale, c'est-à-dire avec un débit accéléré. Ce phénomène pourrait être interprété comme une substitution de la réponse gestuelle et voco-

verbale non formulée par le jeune, qui aurait une double fonction : maintenir le flux interactionnel et fournir une échoïstation empathique (Cosnier, 1988). Il s'agit là de pistes exploratoires qui interrogent sur le statut de ces énoncés mais également la manière de les coder sur le plan prosodique et de les comparer avec celles produites avec des enfants ayant une trajectoire typique.

Mots clés : polyhandicap, interaction, répétition, dialogue, prosodie

Références

BERNICOT, Josie, Anne SALAZAR-ORVIG, et Edy VENEZIANO. 2006. « Les reprises : dialogue, formes, fonctions et ontogenèse ». *La linguistique* 42 : 29-49.

BOCÉRIAN, Christine, Emmanuelle CANUT, Michel MUSIOL, et Virginie ANDRÉ. 2010. « Quel usage les adultes font-ils de la répétition ? Comparaison entre un corpus d'interactions verbales adultes/adolescents polyhandicapés et adultes/jeunes enfants ». *Actes du 2ème Congrès Mondial de Linguistique Française*, publié en ligne : DOI: <https://doi.org/10.1051/cmlf/2010138>.

COSNIER, Jacques. 1988. « Grands tours et petits tours ». J. Cosnier, N. Gelas & C. Kerbrat-Orecchioni (eds) : *Echanges sur la conversation*. Paris : Editions CNRS. 175-184.

HOSTYN, Ine, & Bea MAES. 2009. « Interaction between persons with profound intellectual and multiple disabilities and their partners: a literature review ». *Journal of intellectual & developmental disability* 34(4) : 296-312.

TOUBERT-DUFFORT, Danièle, Esther ATLAN, Hervé BENOIT, Nathalie LEWI-DUMONT, Minna PUUSTINEN, et Rafika ZEBDI. 2018. *Conditions d'accès aux apprentissages des jeunes polyhandicapés en établissements médico-sociaux - de l'évaluation des potentiels cognitifs à la mise en œuvre de leur scolarisation*. Rapport de recherche, INSHEA, Suresnes.

Notice bio-bibliographique

Karine Martel est maîtresse de conférences en psychologie du développement à l'Institut national supérieur de formation et de recherche pour l'éducation inclusive (INSEI), et membre du Groupe de recherche sur le handicap, l'accessibilité, les pratiques éducatives et scolaires (Grhapes - EA 7287). Elle a été formée à l'Université de Nanterre et a soutenu sa thèse de doctorat en psycholinguistique à l'Université de Caen Normandie en 2001. Elle a également validé un diplôme universitaire "Autisme(s) : approche intégrative des troubles envahissants du développement" à la faculté de médecine de Caen, en 2010. Ses recherches portent les processus d'acquisition du langage, le rôle de la prosodie dans les interactions verbales en contexte typique et atypique, parents-jeunes enfants et enseignants-élèves, et les troubles du langage oral notamment dans le cadre de la formation des professionnels de la petite enfance et de l'éducation. Elle s'intéresse en particulier à la manière dont les interlocuteurs utilisent les ressources vocales pour co-construire du sens et structurer leur discours autour de thématiques partagées.

ANNOTATION PROSODIQUE DANS L'ORAL SPONTANÉ



Philippe MARTIN

Université Paris Cité – LLF, UFRL

philippe.martin@u-paris.fr

L'intonation dans la communication orale est essentielle. Elle permet de révéler (ou de simuler) l'émotion et l'origine socio géographique du locuteur, mais surtout, en plus d'indiquer la modalité de la phrase (déclarative, interrogative et ses variantes), elle segmente et structure la parole en unités prosodiques minimales, les groupes accentuels. Ces groupes accentuels, ne comportant qu'une seule syllabe accentuée (non-emphatique), en position finale en français, sont assemblés en plusieurs niveaux, pour constituer la structure prosodique. Cette structure permet à l'auditeur d'appréhender progressivement le sens de l'énoncé par une segmentation du flux de parole, rendue essentielle par la limite temporelle de la mémoire à court terme dans la parole continue, limitée à quelque 2 ou 3 secondes.

Tout système d'annotation des données prosodiques présuppose le choix d'un modèle théorique, ici différent du modèle autosegmental métrique dominant, dans lequel la structure prosodique résulterait d'un mapping à partir de la structure morphosyntaxique de la phrase. Il s'agit au contraire d'envisager une structure prosodique autonome, mise en place par le locuteur par sections successives avant même l'insertion du texte (y compris dans la lecture), et dont les marques prosodiques satisfont aux conditions nécessaires et suffisantes pour indiquer sans ambiguïté les regroupements des groupes accentuels en syntagmes intonatifs. Les groupes accentuels apparaissent alors comme autant d'unités mémorisées de « prêt-à-parler ».

Dans ce modèle, les marques prosodiques, instanciées en français par des contours mélodiques à l'endroit des voyelles des syllabes accentuées, sont classées en a) montant et descendant, b) au-dessus ou en dessous du seuil de glissando (donc perçue comme un changement mélodique ou un ton statique), et c) atteignant le niveau mélodique le plus bas (cas déclaratif) ou le plus haut (cas interrogatif) parmi les autres contours alignés sur les voyelles accentuées de la phrase.

On présentera un logiciel dédié à l'annotation prosodique basé sur ce modèle, logiciel optimisé pour l'analyse des courbes mélodiques d'enregistrements éventuellement dégradés (écho, chevauchement de voix, accompagnement musical, etc.). Ce logiciel est pourvu de nombreuses fonctions facilitant l'annotation, telles que la segmentation automatique du texte en transcription phonétique API, un concordancier intégré pour la recherche d'occurrences lexicales ou syntaxiques dans de grands corpus, le classement automatique en contours mélodiques, l'affichage dynamique des structures prosodiques à partir des contours, etc. On peut ainsi visuellement mieux se rendre

compte de l'assemblage dynamique fait par l'auditeur pour avoir progressivement l'accès au sens de la phrase, à partir des balises temporelles instanciées par les contours mélodiques.

Les évènements prosodiques et leur représentation, ainsi que la grammaire de dépendance déterminant la structure prosodique, peuvent être définies par l'utilisateur, et toutes les données exportées selon divers formats (Excel, csv, etc.). Il est donc possible d'utiliser d'autres modèles, par exemple utilisant la notation ToBI et une grammaire basée sur la morphosyntaxe du texte.

On montrera des exemples d'annotation de dialogues spontanés en français et en italien, illustrant des classes d'évènement prosodiques différents, mises en œuvre par des grammaires spécifiques, sur les données extraites des corpus de parole spontanée Orfeo, Ofrom et C-Oral-Rom.

Mots clés : intonation, groupes accentuels, structure prosodique, contours mélodiques, ToBI

Références

C-Oral-Rom : <http://www.elda.org/en/proj/coralrom.html>

Orfeo : <http://www.projet-orfeo.fr/>

Ofrom : <http://ofrom.unine.ch>

Notice bio-bibliographique

Philippe Martin est professeur émérite de l'université Paris Cité et de l'Université de Toronto. Titulaire d'un doctorat ès Sciences et d'un doctorat en Linguistique, il a enseigné aux Universités de Toronto, d'Aix-en-Provence et au Département de Linguistique de l'Université Paris Diderot, dont il a été directeur de 2004 à 2012. Il a publié de nombreux articles sur la phonologie et la phonétique de l'intonation dans les langues romanes ainsi que sur l'analyse acoustique de la parole. Ses derniers livres sont *Speech Acoustic Analysis* (2021), *Analyse acoustique de la parole* (2020), *Intonation, structure prosodique et ondes cérébrales* (2018), et *The Structure of Spoken Language* (2015). Il a également développé des dispositifs matériels et logiciels pour l'analyse prosodique de grands corpus oraux (WinPitch). Il travaille actuellement sur le traitement neuronal de la structure prosodique dans la parole, et a reçu en 2018 le titre de Docteur Honoris Causa de l'Université Charles, à Prague.

HOW EFFECTIVE ARE AUTOMATIC ANNOTATION TOOLS FOR INVESTIGATING VOCALIC CHANGE IN AUSTRALIAN ENGLISH? AN ANALYSIS OF CONVERSATIONAL SPEECH.



Erwanne MAS

University of Toulouse-Jean Jaurès – CLESCO

erwanne.mas@univ-tlse2.fr

Though annotation of oral data can be time-consuming, phonetic transcriptions can be facilitated with automatic tools. Two automatic tools, one trained on Australian English and New Zealand English authentic spoken data, LaBB-CAT (Fromont & Hay, 2012), and the other not specifically trained on Australian English data, the Montreal Forced Aligner (McAuliffe et al., 2017), were used to transcribe segments of semi-spontaneous conversations between Australian English speakers. Oral data was sampled from conversational speech recorded with a Zoom-H4n Pro digital recorder, set at a resolution of 44.1 kHz/16 bits. Around 10 minutes of speech for 40 speakers was analysed to see the impact of automation on orthographic and phonetic transcription of Australian English oral data. We present how the two automatic tools performed on our data and examine where and why they did not manage to provide an accurate segmentation of the oral data. Oral segments involving multiple speakers within a given segment were compared with segments comprising only one speaker. As automatic tools do not always manage to provide accurate transcriptions of oral data, we show how segmenting spoken data into speech units can help automatic tools to process speech, though this does not necessarily solve the problem and rather leads to new questions related to how automatic annotators manage overlapping voices when compared to human annotators. Australian English indeed presents specific challenges for annotators because of its status as a variety of English and because it is not much used to train automatic automation tools compared to Standard British English or Standard American English. As most automatic tools have not been trained using Australian English spoken data, we show how the performance of annotation tools leads to questions regarding phonetic transcriptions since annotation tools do not always correctly identify phonetic features, and more specifically diphthongs. Australian English presents challenges in phonetic analyses due to regional variation involving vocalic features (Cox & Palethorpe, 2019) and questions linked to how well automatic annotation tools perform on spontaneous speech indeed influence our way of analysing vocalic phonemes and evaluating phonemic contrast in Australian English.

Key words : sociolinguistics, Australian English, phonetics, annotation tools

References

- COX, Felicity, & Sallyanne PALETHORPE. 2019. « Vowel variation in a standard context across four major Australian cities ». Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019 : 577–581. Australasian Speech Science and Technology Association.
- FROMONT, Robert, & Jennifer HAY. 2012. « LaBB-CAT: An Annotation Store ». Proceedings of the Australasian Language Technology Association Workshop 2012 : 113–117. Dunedin, New Zealand.
- GONZALEZ, Simon, James GRAMA, & Catherine E. TRAVIS. 2020. « Comparing the performance of forced aligners used in sociophonetic research ». Linguistics Vanguard 6(1), 20190058. <https://doi.org/10.1515/lingvan-2019-0058>.
- MAS, Erwanne, & Anne PRZEWOZNY-DESRIAUX. 2023. Phonology of Contemporary English: Australia. Cognition, Langues, Langage, Ergonomie; Laboratoire Parole et Langage; Groupe de Recherche sur les Grammaires - Équipe d'Accueil 370; Centre de linguistique interlangues, lexicologie, linguistique anglaise et de corpus - Atelier de Recherche sur la Parole. <https://doi.org/10.34847/COCOON.AB9E1776-BD3F-44E4-9E17-76BD3FB4E4BE>.
- MCAULIFFE, Michael, Michaela SOCOLOF, Sarah MIHUC, Michael WAGNER, & Morgan SONDEREGGER. 2017. « Montreal Forced Aligner: Trainable text-speech alignment using Kaldi ». Proceedings of the 18th Conference of the International Speech Communication Association.
- PETERS, Pam, Peter COLLINS, & Adam SMITH. (Eds.) 2009. Comparative studies in Australian and New Zealand English: grammar and beyond. (Varieties of English around the world). John Benjamins.

Notice bio-bibliographique

Ancienne élève de l'ENS de Lyon au sein du département des langues, littératures et civilisations, **Erwanne MAS** est actuellement en doctorat en linguistique anglaise à l'Université Toulouse Jean Jaurès. Sa thèse de doctorat porte sur les changements vocaliques de l'anglais australien.

LA LINGUISTIQUE DE L'ORAL À TRAVERS LES LANGUES SUR LES PROBLÈMES DE TRANSCRIPTION DES LANGUES PARLÉES PEU DOTÉES



Samira MOUKRIM

Université SMBA de Fès – Discredor (CALL)

Université d'Orléans – LLL (UMR 7270)

samiramoukrim@yahoo.fr

Une langue n'a d'existence pour la recherche qu'en fonction de la qualité de ressources préparées pour son exploitation. L'avènement, ces dernières années, de la linguistique dite de « corpus », l'élaboration de « bonnes pratiques » par une communauté de chercheurs qui travaillent sur les corpus et l'utilisation des technologies informatiques actuelles, permettent d'apporter des solutions innovantes pour la description des langues. Si les grandes langues véhiculaires, notamment européennes, ont très largement bénéficié de ces innovations, les langues peu dotées sont restées à l'écart de ce mouvement et ont accusé un retard important.

Outre des raisons historiques et symboliques, puisqu'il s'agit essentiellement de variétés parlées vernaculaires, considérées jusqu'à très récemment comme peu dignes d'intérêt, les raisons de ce retard sont multiples. On peut évoquer ici les difficultés liées à l'édition des données orales, et les questions scientifiques et techniques sous-jacentes qu'elle soulève. La constitution, l'exploitation (notamment sous la forme d'investigation instrumentée) et la mise en commun de grands corpus de ces langues restent un objectif encore à atteindre, dont la première étape consiste à structurer une communauté de chercheurs aux pratiques sinon homogènes, du moins interoperables.

L'un des chantiers qu'une communauté structurée de chercheurs devrait rouvrir est celui de la transcription qui constitue un verrou à la documentation et l'exploitation outillée des données orales. Si, depuis une décennie, des projets visant à produire des ressources et des outils pour les langues peu dotées commencent à connaître un regain d'intérêt, la profusion de modes de transcription, de formats et de conventions rend difficile leur exploitation.

L'objectif de cette communication est de partager mon expérience et mes acquis dans le domaine de la constitution et l'exploitation de corpus oraux en mettant l'accent sur les problèmes de transcription des langues peu dotées.

Le corpus d'étude a été constitué dans le cadre du programme Langues en Contact à Orléans (LCO), module du Projet Enquêtes sociolinguistiques à Orléans 2 (ESLO2), piloté par le Laboratoire Ligérien de Linguistique de l'Université d'Orléans (LLL). Le programme LCO avait pour objectif de dresser un portrait du multilinguisme dans la ville d'Orléans. Il se propose : i) de répertorier les pratiques linguistiques et culturelles des immigrés aux moyens d'enquêtes sur le terrain ; ii) d'analyser la vie de ces langues. Quels sont leurs modes de transmission ? Où sont-elles parlées,

par qui, dans quelles circonstances ? iii) de réaliser des enregistrements des différentes langues telles que parlées à Orléans pour mettre à disposition de la communauté scientifique et pourquoi pas de façon plus large.

Un sous-corpus (du français, de l'arabe dialectal et du berbère tamazight) a été constitué auprès de locuteurs arabophones et amazighophones (berbérophones) résidant à Orléans. Pour améliorer la représentativité du corpus, nous avons diversifié les situations enregistrées ainsi que les catégories de locuteurs en différenciant sociologiquement les témoins par l'âge, le sexe, le niveau scolaire, la profession et les langues parlées.

Nous avons constitué un corpus de « données situées » qui contient, en plus des données primaires (les enregistrements de la parole), une riche documentation sur ces données et sur leur contexte de production. Nous avons tenu également à expliciter notre démarche, à documenter les conditions de constitution du corpus ainsi que nos choix théoriques et techniques. Toutes ces informations sont d'une grande importance pour rendre le corpus disponible.

La qualité de la constitution du corpus est un facteur de la relation de confiance établie entre nous et les enquêtés. Notre relation privilégiée avec les témoins vient du fait que nous partageons avec eux un certain nombre de traits : même origine (pays d'origine), même langue(s), même situation en France (appartenance à la population immigrée), etc.). Nous sommes membre à part entière de la culture vernaculaire. Nous la connaissons de l'intérieur et nous la comprenons profondément. Cela permet, selon Labov (1978 : 10) de « réussir une percée plus profonde ».

La collecte des données s'est effectuée sous forme d'enregistrements. Ces données sonores brutes ne peuvent pas être analysées sans un travail préalable de transcription et de segmentation. Le processus de transcription soulève un certain nombre d'enjeux théoriques, méthodologiques et pratiques. Transcrire ne peut se résumer à un simple exercice de reproduction des séquences perçues. Les différents choix, notamment celui de l'orthographe peuvent avoir une incidence sur la lisibilité, l'analyse et l'automatisation des données.

Transcrire des langues à tradition essentiellement orale soulève des problèmes spécifiques, surtout si l'on entend produire un corpus partageable. En plus des difficultés que pose toute transcription de l'oral, d'autres interrogations, d'une autre nature, surgissent lors de la phase préparatoire à la transcription des langues à tradition orale et pour lesquelles il existe une concurrence entre plusieurs alphabets et donc plusieurs traditions orthographiques, des problèmes liés à la fois au système graphique, au mode, aux conventions et aux outils de transcription à adopter...

L'hétérogénéité des trois corpus du français, de l'arabe marocain et du berbère tamazight se manifeste tant au niveau de la langue, différente d'un corpus à l'autre, qu'au niveau de sa codification graphique et du degré de son informatisation. Si le français dispose d'une tradition écrite ancienne et bien établie (standard), ce n'est le cas ni du berbère tamazight ni de l'arabe marocain. Le berbère tamazight est une langue à tradition orale, sans tradition écrite standard ni dialecte de référence. L'arabe marocain est une langue à tradition orale, pour laquelle il existe certes un standard, l'arabe classique, auquel il est apparenté, mais dont le système linguistique est très différent.

Le choix d'une notation de l'arabe dialectal ou du berbère tamazight implique certains a priori et la sélection d'un mode de transcription (phonétique, phonologique ou orthographique) sous son apparence technique infère des orientations dont la signification sociolinguistique et épistémologique doit être interrogée. Nous allons voir, à travers cette étude, en quoi la différence des corpus implique une spécification des méthodes et outils, en particulier si les outils et instruments utilisés pour les corpus oraux de langues à grande diffusion (notamment celui du français) s'ajustent aux langues à tradition orale.

Références

- BAUDE, Olivier. 2006. Corpus oraux, Guide des bonnes pratiques. C.N.R.S. éditions et PUO.
- BILGER, Mireille. 2008. « Les enjeux des choix orthographiques », Bilger, Mireille (éd.) Données orales – Les enjeux de la transcription. Perpignan. PUP : 248-257 .
- BLANCHE-BENVENISTE, Claire & Colette JEANJEAN. 1987. Le français parlé – Édition et transcription. Paris: Didier-Érudition.
- BLANCHE-BENVENISTE, Claire. 1990. Le français parlé – Études grammaticales. Paris : CNRS. dans le dialogue oral spontané contraint : le cas du Contrôle Aérien, Thèse de doctorat, Université Paul.
- CHAKER, Salem. 1996. « Propositions pour la notation usuelle à base latine du berbère ». Atelier du 24-25 juin 1996, INALCO-CRB. Synthèse des travaux.
- CHAKER, Salem. 2002. « Variation dialectale et codification graphique en berbère. Une notation usuelle pan-berbère est-elle possible ? », D. Caubet, S. Chaker, J. Sibille (éds), Codification des langues de France. L'Harmattan : Paris. 331-340.
- GADET, Françoise. 2000. « Derrière les problèmes méthodologiques du recueil des données ». M. Bilger (dir.) Linguistique sur corpus. Presses Universitaires de Perpignan.
- GADET, Françoise. 2008. « L'oreille et l'œil à l'écoute du social ». Bilger, Mireille (éd.) Données orales: les enjeux de la transcription. Perpignan : PUP. 35-47.
- CANDEA, Maria. 2000. Contribution à l'étude des pauses silencieuses et des phénomènes dits « d'hésitation" » en français oral spontané. Thèse d'État, Université Paris III (Sorbonne Nouvelle).
- CHANET, Catherine. 2004. Fréquence des marqueurs discursifs en français parlé : quelques problèmes de méthodologie. Recherches sur le français parlé 18 : 83-105.
- CONSTANT, Matthieu, & Anne, DISTER. 2012. « Les disfluences dans les mots composés ». Manuscrit auteur, publié dans « Journées sur l'Analyse des Données Textuelles (JADT'12) ». Belgique.
- DISTER, Anne. 2007. De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelles orales Valibel. Thèse de doctorat. Université de Louvain.
- HABERT, Benoît, Adeline NAZARENKO & André SALEM. 1997. Les linguistiques de corpus. Paris : A. Colin.
- HABERT, Benoît. 2000. « Des corpus représentatifs : de quoi, pour quoi, comment ? ». M. Bilger (dir.), Linguistique sur corpus. Presses universitaires de Perpignan.

HENRY, Sandrine. 2002. Étude des répétitions en français parlé spontané pour les technologies de la parole. Actes de la 6ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues. Nancy (France). 467-476.

HENRY Sandrine, Estelle CAMPIONE & Jean VÉRONIS. 2004. « Répétitions et pauses (silencieuses et remplies) en français spontané ». JEP'04. Fès, Maroc. 261-264.

JACQUES, Marie-Paule. 2005. « Pourquoi une linguistique de corpus ? ». La linguistique de corpus, G. Williams éd. Rennes : Presses universitaires de Rennes (Rivages linguistiques). 21-30.

KURDI, Mohamed-Zakaria. 2003. Contribution à l'analyse du langage oral spontané. Thèse de doctorat, Université J. Fourier. Grenoble, France.

PALLAUD, Bertille. 2002. Les amorces de mots comme faits autonymiques en langage oral. Recherches sur le Français parlé 17 : 79-102.

PIU, Marie. 2006. Annotation des disfluences dans les corpus oraux. Mémoire de Master. Université de Provence, Aix-en-Provence.

VÉRONIS, Jean. 1998. Annotation automatique de corpus : état de la technique. Colloque International « Questions de méthode dans la linguistique de corpus ». Perpignan.

Notice bio-bibliographique

Samira MOUKRIM est professeure des Universités en Linguistique et Sciences du Langage à l'Université SMBA de Fès. Elle est chef du département de Langue et Littérature Amazighes à la Faculté des Lettres Saïs-Fès, chercheuse associée au LLL - Université d'Orléans et membre du Dipralang - Université Paul-Valéry Montpellier 3.

Elle a obtenu son doctorat en Linguistique et Sciences du Langage de l'Université d'Orléans, sous la direction des professeurs Gabriel BERGOUNIOUX et Lotfi ABOUDA, le 10 décembre 2010. Intitulé de la thèse : Morphosyntaxe et sémantique du « présent » : étude contrastive à partir de corpus oraux (ESLO/LCO) : Arabe marocain, Berbère tamazight et Français. Qualification : sections 7 et 15.

ON THE TWO REALMS OF SPOKEN SPANISH: A STUDY BASED ON DISCOURSE MARKERS POSITIONS



Adrià PARDO LLIBRER

University of Salamanca

pardolibrer@usal.es

This presentation shows how Spanish discourse markers are specialized in two realms of spoken language based on their functions and positions. In my approach, discourse markers present textual, modal, or interpersonal functions; these functions are described according to their scope and position regarding a discourse unit (Briz Gómez and Pons Bordería, 2010; Pons Bordería, 2018; Pons Bordería, Pardo Llibrer, and Alemany Martínez, 2023). To achieve this description, I follow the Val.Es.Co model of discourse segmentation in units and positions (Val.Es.Co. Research Group, 2014; Pons, 2016, 2022). My study applies this segmentation model to all examples from the DPDE – or Diccionario de partículas discursivas del español ('Dictionary of Spanish Discourse Particles'; Briz, Pons, and Portolés, 2008). This dictionary describes around a hundred discourse markers, using examples from colloquial conversations corpora (Briz Gómez and Val.Es.Co. Research Group, 2003; Pons Bordería, online) and written texts that reproduce spoken language (López-Serena and Sáez Rivera, 2018).

Segmenting the DPDE examples offers sound linguistic data to classify types of discourse markers based on the variables function, scope, position, and genre. An initial descriptive quantification of the data helps to identify those specific functions characteristic of written texts. After this genre filter, I applied a binomial distribution to those markers of spoken language, which significantly associates discourse markers exhibiting textual and modal functions with the initial position of speech; by their part, interpersonal discourse functions associate with the final position. I finally conducted an agglomerative-hierarchical cluster analysis, underlining the variable position as the central trait to distinguish discursive functions. Thus, the findings point to two main clusters of discourse markers and reveal a division into two realms in spoken Spanish: an (unmarked) oral realm, involving the initial position as the default position of discourse markers, and a marked oral realm, involving those items in final or independent positions.

Key words : discourse markers, functions, positions, oral realms, marked orality

References

BRIZ GÓMEZ, Antonio & Salvador PONS. 2010. "Unidades, marcadores discursivos y posición." In *Los estudios sobre marcadores del discurso en español*, ed. by Óscar Loureda, and Esperanza Acín-Villa, 327-358. Madrid: Arco-Libros.

BRIZ GÓMEZ, Antonio, Salvador PONS, & José PORTOLÉS. 2008. *Diccionario de partículas discursivas del español*. Online available: <www.dpde.es>.

BRIZ GÓMEZ, Antonio & Val.Es.Co. Research Group. 2003. *Corpus de conversaciones coloquiales*. Madrid: Arco-Libros.

LÓPEZ-SERENA, Araceli & Daniel M. SÁEZ RIVERA. 2018. "Procedimientos de mimesis de la oralidad en el teatro español del siglo XVIII." *Estudios Humanísticos: Filología* 40: 235-273.

PONS BORDERÍA, Salvador, Adrià PARDO LLIBRER, & Amparo ALEMANY. 2023. *La marcación discursiva en español: descripción y análisis conjunto desde el DPDE*. Sevilla: EUS.

PONS BORDERÍA, Salvador (dir.): *Corpus Val.Es.Co 3.0*. <www.valesco.es>.

PONS BORDERÍA, Salvador. 2022. *Creación y análisis de corpus orales: saberes prácticos y reflexiones teóricas*. Berlin: Peter Lang.

PONS BORDERÍA, Salvador. 2018. "Paths of Grammaticalization: Beyond the LP/RP Debate." In *Beyond Grammaticalization and Discourse Markers. New Issues in the Study of Language Change*, ed. by Maj-Britt Mosegaard Hansen, Kerstin Fischer, and Anne Barron, Leiden/Boston: Brill. 334–383.

PONS BORDERÍA, Salvador. 2016. "Cómo dividir una conversación en actos y subactos". In *Oralidad y análisis del discurso: homenaje a Luis Cortés Rodríguez*, ed. by Antonio M. Bañón, M. del Mar Espejo, Bárbara Herrero, and Juan L. López. Almería: Universidad de Almería. 545-566.

Val.Es.Co. Research Group. 2014. "Las unidades del discurso oral. La propuesta Val.Es.Co. de segmentación de la conversación (coloquial)." *Estudios de Lingüística del Español* 35: 13-73.

Biography

Adrià Pardo Llibrer is assistant professor at the University of Salamanca. His research focuses on the study of discourse markers in spoken Spanish, applying segmentation models. He is also interested in the diachronic study of the meaning change of pragmatic devices.

TRANSCRIRE OU NE PAS TRANSCRIRE, TELLE EST LA QUESTION : LE CHOIX DE TRANSCRIPTION DE L'ORAL DANS LE CADRE D'UN PROJET EUROPÉEN



Liubov PATRUKHINA

Université Toulouse Jean Jaurès – CREG

liubov.patrukhina@univ-tlse2.fr

Les cadres théoriques, mais aussi les cultures scientifiques du pays ou de la langue étudiée, déterminent fortement la façon dont un chercheur transcrirait un corpus oral. Les choix de transcription peuvent de ce fait nous éclairer sur le positionnement théorique adopté face à l'analyse de l'oral, sans que ce choix fasse forcément l'objet d'une réflexion théorique préalable de la part du transcripteur. Ainsi, la tradition qui s'est progressivement installée dans les pays de langue allemande au tournant du XXI^{ème} siècle impose une façon de transcrire assez complexe, héritée des recherches en phonologie et nécessitant une prise en main de la technique (cf. Selting et al. 2009). Les grands corpus oraux de l'allemand faisant partie du DGD (Datenbank Gesprochenes Deutsch = « Banque de données de l'allemand parlé ») adoptent notamment ces conventions. Au contraire, les choix de transcription principalement orthographique, utilisés par la majorité des grands corpus de français parlé actuels (ESLO, CLAPI, PFC, CFPP2000, cf. Calinon et al. 2019), reflètent la volonté méthodologique de mettre en place une transcription plus accessible (cf. Blanche-Benveniste 1990).

La participation aux projets européens autour d'un corpus commun permet de se questionner sur ses propres choix de transcription de l'oral, dans le souci de trouver la manière de transcrire la plus adaptée à l'objet d'étude, mais aussi au profil des chercheurs-participants. Dans le cadre de cette communication, j'ai choisi comme objet d'analyse purement qualitative deux vidéos enregistrées via Zoom qui montrent des enseignants-chercheurs et/ou étudiants dans un contexte académique. Toutes les deux étant des exemples d'une communication par écran interposé, la vidéo en allemand présente une discussion sur la mise en place d'un projet de recherche, alors que la vidéo en français retrace le déroulement d'un séminaire. Si le nombre de participants, le cadre communicationnel et les thématiques abordées restent comparables, la proportion des locuteurs non-natifs et le lien (hiérarchique) entre les participants sont différents.

À l'appui de ces vidéos, je souhaiterais explorer les questions plus globales, pertinentes pour ce corpus trilingue (allemand, français, italien). Pourrait-on envisager une transcription « intermédiaire » qui conviendrait à tous les participants ou serait-il plus pertinent d'adopter une transcription multiple en fonction des besoins de chaque chercheur-participant du projet ? Quels paramètres supplémentaires convient-il de prendre en compte lors de l'analyse des interactions par écran interposé par rapport aux autres types d'interactions filmées (Ibdelkaïd 2019) ? L'analyse de l'aspect multimodal des interactions, est-il incontournable (cf. Baldauf-Quilliatre & Colón de

Carvajal 2022) ? Lors d'une analyse plus fine : faudrait-il prendre en compte les particularités de l'oral des locuteurs non-natifs ? Quel degré de standardisation convient-il d'adopter ? L'enjeu de ma communication est donc de proposer un modèle de (non-)transcription de l'oral et des activités non-verbales qui permettrait aux chercheurs-participants de ce projet de travailler efficacement ensemble sur un corpus commun.

Mots clés : transcription de corpus oraux, communication par écran interposé, multimodalité, cultures scientifiques

Références

BALDAUF-QUILLIATRE, Heike, & Isabel COLÓN DE CARVAJAL. 2022. « Multimodal practice of participation in a complex and dynamic framework ». Lund, K. et al. (Éds.), *Language is a complex adaptive system: Explorations and evidence*. Berlin : Language Science Press. 141-162.

BLANCHE-BENVENISTE, Claire et al. 1990. *Le Français parlé : Études grammaticales*. Paris : CNRS.

CALINON, Anne-Sophie, Hamma BADREDDINE, Katja PLOOG, et Marie SKROVEC. 2019. *Linguistique interactionnelle, grammaire de l'oral et didactique du français*. Besançon : Presses universitaires de Franche-Comté.


IBNELKAÏD, Samira. 2019. « Chapitre 10. L'agentivité multimodale en interaction par écran, entre sujet et tekhnê ». Mazur-Palandre, A. & Colón de Carvajal, I., *Multimodalité du langage dans les interactions et l'acquisition*. Grenoble : UGA Éditions. 281-315.

SELTING, Margret, et al. 2009. « Gesprächsanalytisches Transkriptionssystem 2 (GAT2) ». *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion* vol. 10 : 353-402.

Notice bio-bibliographique

Originaire de Saint-Pétersbourg, **Liubov Patrukhina** a fait sa thèse à l'Université Sorbonne Nouvelle. Ses intérêts de recherche touchent à la fois à la linguistique contrastive et à la didactique des langues, avec une attention particulière portée aux particules modales allemandes. Depuis septembre 2020, elle occupe un poste de maître de conférences en linguistique allemande au sein de l'Université Toulouse – Jean Jaurès (LLCE DLE Section d'allemand / EA 4151 CREG).

UNITÉS DISCURSIVES, CHUNKS SYNTAXIQUES ET SYNTAGMES ACCENTUELS EN FRANÇAIS CONVERSATIONNEL



Laurent PRÉVOT

Aix-Marseille Université, CNRS – LPL

CNRS & MEAE – CEFC

laurent.prevot@univ-amu.fr

Roxane BERTRAND

Aix-Marseille Université, CNRS – LPL

roxane.bertrand@univ-amu.fr

Bien que la segmentation automatique en unités discursives élémentaires (EDU) (Asher & Lascarides, 2003) soit parfois considérée comme une tâche résolue, cette appréciation ne peut être défendue que pour certains genres de l'écrit et dans le contexte de langues richement dotées en outils et modèles de langues (LLMs). Dès que l'objet d'intérêt s'éloigne de ce standard où ponctuation et segmenteurs phrastiques ne sont pas disponibles ou peu opérant la question devient plus délicate. Or certains travaux (Prévot et al., 2023 ; Prévot & Wang, 2024), montrent que l'adaptation (fine-tuning) de LLMs sur de l'oral spontané et sur des langues moins dotées (que l'anglais ou le français) fournit des résultats proches de l'accord entre annotateurs humains, ce qui permet de disposer, en grande quantité, d'unités de discours du genre conversationnel pour un nombre de langues croissant.

Ce travail nous permet d'approcher les unités discursives des corpus conversationnels en nous fondant sur leur contribution sémantique et pragmatique et ce qu'elle que soit leur forme : phrase complète, proposition indépendante, énoncé non-phrase (Fernandez et al., 2007). Nous étudions la composition de ces unités en termes de chunks (Abney, 1991) et de syntagmes accentuels (Jun & Fougeron 2000), lesquels sont supposés être fortement liés (Abney, 1991). Ces éléments constituent selon nous des unités de traitement linguistique cruciales pour les locuteurs. Ce travail aborde deux questions principales. Premièrement, nous vérifions l'hypothèse de l'association entre les unités de base des traitements syntaxique et prosodique, ce qui à notre connaissance n'a pas été fait systématiquement en conversation. Ensuite, nous avançons qu'en conversation la structure interne des unités discursives (en termes de chunks syntaxiques ou de syntagmes accentuels) possède une grammaire relativement simple dont la clarté a été offusquée par des normes d'analyses, des lexiques et grammaires établies et pensées pour les genres écrits. Cette remarque n'est pas nouvelle (Blanche-Benveniste, 1996) mais nous proposons un cadre d'analyse systématique et quantitatif renouvelant la réflexion sur ce sujet via les questions des disfluences, marqueurs discursifs et énoncés non-phrase.

Méthodologiquement, nous avons tout d'abord segmenté en unités discursives des corpus de parole conversationnelle à l'aide de segmenteurs que nous avons adapté à ce genre (Prévot et al, 2023) qui ont été entraînés sur des annotations manuelles réalisées en amont (Prévot et al, 2021). Nous disposons par ailleurs d'annotations prosodiques à deux niveaux : syntagmes intonatifs et syntagmes accentuels. Nous menons actuellement une campagne d'annotation en chunks selon une logique proche de (Eshkol-Taravela et al, 2020). Nous observons tout d'abord que plus de 90% des DU sont composés de 7 chunks ou moins. Dans notre analyse, certaines de nos catégories de chunks correspondent à des unités fonctionnelles pouvant correspondre à des lexicalisations, typiquement de séquences de mots de fonction comme les structures présentationnelles.

Dans la présentation, nous reviendrons rapidement sur ces différentes segmentations avant d'approfondir les associations entre prosodie, syntaxe et discours.

Mots clés : segmentation, conversation, unité discursive, chunk, syntagme accentuel, français

Références

- ABNEY, Steven P. 1992. Parsing by chunks. *Principle-based parsing: Computation and Psycholinguistics*. 257-278.
- ASHER, Nicholas, & Alex LASCARIDES. 2003. *Logics of conversation*. Cambridge University Press.
- BLANCHE-BENVENISTE, Claire. 1996. Trois remarques sur l'ordre des mots dans la langue parlée. *Langue française*. 109-117.
- ESHKOL-TARAVELLA, Iris, Mariame MAAROUF, Flora BADIN, Marie SKROVEC, & Isabelle TELLIER. 2020. « Chunk Different Kind of Spoken Discourse: Challenges for Machine Learning ». *Language Resources and Evaluation Conference* : 5164-5168.
- FERNÁNDEZ, Raquel, Jonathan GINZBURG, & Shalom LAPPIN. 2007. « Classifying non-sentential utterances in dialogue: A machine learning approach ». *Computational Linguistics* 33(3) : 397-427.
- JUN, Sun-Ah, & Cecile FOUGERON. 2000. « A Phonological model of French intonation ». Antonis Botinis (ed.) *Intonation: Analysis, Modeling and Technology*, Kluwer Academic Publishers: Dordrecht. 209-242.
- PRÉVOT, Laurent, Roxane BERTRAND, & Stéphane RAUZY. 2021. « Investigating disfluencies contribution to discourse-prosody mismatches in french conversations ». *The 10th Workshop on Disfluency in Spontaneous Speech*.
- PRÉVOT, Laurent, Julie HUNTER, & Philippe MULLER. 2023. « Comparing Methods for Segmenting Elementary Discourse Units in a French Conversational Corpus ». *24th Nordic Conference on Computational Linguistics (NoDaLiDa 2023)*.
- PRÉVOT, Laurent, & Sheng-Fu WANG. (submitted) *Experimenting with Discourse Segmentation of Taiwan Southern Min Spontaneous Speech*.

Notice bio-bibliographique

Laurent Prévot est professeur de Sciences du Langage à l'université Aix Marseille (AMU) et chercheur au Laboratoire Parole et Langage (UMR 7309). Il a soutenu une thèse d'informatique (2004) à l'Université Paul Sabatier (Toulouse) et une HDR (2014) en Sciences du Langage à Aix-Marseille Université. Il est actuellement en délégation CNRS au CEFC à Taipei (Taiwan). Son doctorat portait sur la modélisation des structures sémantiques et pragmatiques dans les dialogues finalisés. Il a publié de nombreux articles et chapitre ayant utilisant des méthodes de linguistique computationnelle et d'apprentissage automatique pour de nombreux sujets liés, en particulier, au genre oral spontané.

Roxane Bertrand est directrice de recherche au CNRS, au laboratoire Parole et Langage à l'Université Aix-Marseille (AMU). Elle a soutenu sa thèse en Sciences du Langage (1999) et une HDR (2021). Ses travaux portent sur la conversation qui est conçue comme une activité collaborative dans laquelle tous les participants, locuteur et interlocuteur, jouent un rôle crucial dans l'élaboration et le partage du sens. Elle a publié de nombreux articles et chapitres sur des thèmes tels que la convergence interactionnelle ou le feedback, dans lesquels elle conjugue une approche méthodologique inspirée de la Linguistique Interactionnelle et de la Linguistique de Corpus pour rendre compte des liens entre les différentes modalités verbale, vocale et mimo-gestuelle déployées dans l'interaction interindividuelle.

INTRALOCUTIONARY AND INTERLOCUTIONARY COHERENCE



Thierry RUCHOT

Université de Caen Normandie – CRISCO

thierry.ruchot@unicaen.fr

There are two traditions running in parallel in the study of discourse. The first one is more concerned with monological relationships, especially in writing. It studies the introduction, maintenance, change or reintroduction of topics, temporal relations, discourse markers of connection, and the construction of narrative, expository or argumentative sequences. In the same vein but sometimes forgotten or kept apart in the work upon coherence, the study of discourse relations (also called rhetorical relations) is concerned with the relationships of interpretation between discourse units (like elaboration, cause, result, contrast, list, etc.), whether these units are called sentences, statements or other. Those relations may be formally marked or unmarked. Their inventory ranges from some 10 relations to more than 100 (Mann et Thompson 1988; Carlson et Marcu 2001; Taboada 2004; Garrido Medina 2007; Duque 2016; Инькова et Манзотти 2019) The other tradition is concerned with the interpretation of acts performed in interaction, either on the basis of an a priori inventory (Searle 1969; Vanderveken 1990) or through the study of discourse in interaction (Couper-Kuhlen et Selting 2017). My aim is to show the complementarity of these two currents when applied to dialogical oral discourse, through the notion of interact as a unit of interlocutory construction and that of intralocutory acts, playing a role in the construction of the monological units of dialogue. While the interact may be limited to a single utterance, representing a unit of interaction consistent with the interaction in progress it may require a more complex internal construction, sometimes making it possible to move from a reactive interact to an initiative interact, restarting the exchange or changing its course. Some models have already been proposed to take into account various strands of research, the most notable being probably the Geneva modular model of discourse analysis developed by Roulet and colleagues (Roulet et al. 2001). My aim here will be to pursue this line of research by showing that we can propose a more integrative approach, associating some dimensions that are often kept separate. So, the informational structure is rarely used to deal with dialogue, and the same is true for the discourse relations. We shall also recognize that some dimensions are specific to interlocutory relations (appellatives, simple phaticals and evaluative phaticals) while others play a role in Intralocutionary relations (some discourse markers seem to be more specifically intralocutionary). The study will be illustrated by examples of real dialogues in French and Russian.

Références

- CARLSON, Lynn, et Daniel MARCU. 2001. « Discourse tagging reference manual ». ISI Technical Report ISI-TR-545 54 (2001): 56.
- COUPER-KUHLEN, Elizabeth, et Margret SELTING. 2017. Interactional linguistics: study language in social interaction. Cambridge, United Kingdom ; New York: Cambridge University Press.
- DUQUE, Eladio. 2016. Las relaciones de discurso. Cuadernos de lengua española 127. Madrid: Arco Libros, S. L.
- GARRIDO MEDINA, Joaquín. 2007. « Relaciones de discurso ». Pandora: revue d'études hispaniques, no 7: 305-32.
- MANN, William C., et Sandra A. THOMPSON. 1988. « Rhetorical Structure Theory: Toward a functional theory of text organization ». Text - Interdisciplinary Journal for the Study of Discourse 8 (3). <https://doi.org/10.1515/text.1.1988.8.3.243>.
- ROULET, Eddy, Laurent FILLIETTAZ, Anne GROBET, et Marcel BURGER. 2001. Un modèle et un instrument d'analyse de l'organisation du discours. Bern, Suisse.
- SEARLE, John R. 1969. Speech acts: an essay in the philosophy of language. London: Cambridge University Press.
- TABOADA, María Teresa. 2004. Building coherence and cohesion: task-oriented dialogue in English and Spanish. Pragmatics & beyond, new ser. v. 129. Amsterdam: John Benjamins Pub. Co.
- VANDERVEKEN, Daniel. 1990. Meaning and speech acts. Cambridge ; New York: Cambridge University Press.
- ИНЬКОВА, Ольга Юрьевна, et Эмилио МАНЗОТТИ. 2019. Связность текста: мерелогические логико-семантические отношения. Москва: Яск.

Notice bio-bibliographique

Thierry Ruchot est maître de conférences à l'Université de Caen et directeur du laboratoire CRISCO (Centre de Recherches Inter-langues sur la Signification en Contexte). Il enseigne la grammaire et la linguistique russe et intervient aussi dans les formations de Sciences du Langage et orthophonie. Ses recherches portent sur la syntaxe et la sémantique russes, mais aussi sur l'analyse des actes de langage, des types énonciatifs de phrases, des modalités, et des interactions orales en russe. Il consacre aussi une part de sa recherche à l'étude du discours médiatique russe.

SEGMENTING SPONTANEOUS SPOKEN LANGUAGE: AN EXPLORATION OF PROSODY AND PERCEPTUAL STRATEGIES IN ITALIAN AND CHINESE



Valentina SACCONI

University of Florence

valentina.saccone@unifi.it

Lorenzo GREGORI

University of Florence

lorenzo.gregori@unifi.it

Shuai LUO

University of Florence

shuai.luo@unifi.it

Spoken language, a dynamic form of communication, unfolds as continuous speech flow. Despite its continuity, comprehension of oral discourse relies on segmentation, a key tool for unraveling how the mind processes spoken information. This involves systematically dividing speech into units rooted in cognitive mechanisms aiding linguistic comprehension. Theories propose processes such as pattern recognition, guiding the identification of meaningful units within the speech flow based on prosody and pragmatic values (Cresti, 2000; Barbosa & Raso, 2018; Barth-Weingarten, 2016), syntax, semantics, or a combination thereof (among others, Ford et al., 1996; Foster et al., 2000). However, can we rely on and share our perceptions? What features enable the perception of boundaries in speech flow? Understanding how individuals from different linguistic and cultural backgrounds parse natural speech is essential for advancing our comprehension of cross-cultural communication and cognitive processes.

This study investigates the strategies employed by speakers in segmenting spoken language, focusing on Italian and Chinese. It explores the impacts of prosodic features, including intonation variation, intensity changes, pauses and rhythm, and semantico-syntactic aspects of the locutive content in providing the structure for speech comprehension.

To assess the perception of boundaries in the speech flow, a listening test on a panel of 48 participants, evenly divided between Italian and Chinese native speakers, will be conducted. Participants will listen to 25 audio stimuli per language, sourced from natural speech corpora (IT: Lablita Corpus, Cresti et al., 2022a; CN: C-Oral-Cin, Cresti et al., 2022b), balanced in terms of speakers' characteristics (gender, age), interaction type (monologue, dialogue, conversation) and duration (average: 7 sec). The panel will be divided into three groups for three different versions of the audio stimuli: i. original speech segments, ii. textually blurred content with original prosody, and iii. audio with flattened intonation contour retaining the text and the original rhythmic

features. Recordings avoid overlapping between different speakers while maintaining spontaneous speech challenges such as disfluencies, hesitations, and hypo-/co-articulation phenomena.

Participants will actively segment the speech flow, identifying boundaries where they perceive a completion of the minimal stretch of speech that can have a communicative function based on their native language proficiency. Cross-cultural differences are expected, highlighting unique strategies used by Italian and Chinese speakers. Initial findings on 10 recordings, evaluated by 4 participants, reveal lower inter-rater agreement (63% Italian, 58% Chinese) in segmenting flattened voice audio, highlighting the importance of prosody in speech segmentation. Italian annotators achieved 80% agreement on original segments, while Chinese annotators reached 84% on audio with prosody and no text.

This methodology captures the interplay between linguistic characteristics and perceptual strategies employed by individuals from diverse language backgrounds. Comparing results in the three versions of the test, it is possible to disentangle the role of prosodic features, such as intonation contour and rhythm, that lead to discontinuities in the speech, from the semantico-syntactic extent of the locutive content. The research contributes to linguistics and holds implications for language education and intercultural communication.

Key words : segmentation, prosody, perception, spoken language, corpus linguistics, Italian, Chinese

References

- CRESTI, Emanuela. 2000. *Corpus di italiano parlato*. Firenze: Accademia della Crusca.
- CRESTI, Emanuela, Lorenzo GREGORI, Massimo MONEGLIA, Carlota N. MARTINEZ & Alessandro PANUNZI. 2022a. The LABLITA Speech Resources. In Cresti, E. & Moneglia, M. (eds). *Corpora e Studi Linguistici*. Milano:Officinaventuno. 85-108.
- CRESTI, Emanuela, Shuai LUO, Massimo MONEGLIA, & Alessandro PANUNZI. 2022b. The annotation of information structure in spoken Chinese: a pilot study. In Botinis, A. (ed). *Proceedings ExLing 2022: 13th International Conference of Experimental Linguistics*. Paris: Université Paris Cité. 41-44.
- BARBOSA, Plinio & Tommaso RASO. 2018. Spontaneous speech segmentation: Functional and prosodic aspects with applications for automatic segmentation. *Revista de Estudos da Linguagem*, 26(4), 1361–1396.
- BARTH-WEINGARTEN, Dagmar. 2016. Intonation units revisited. *Cesura in talk-in-interaction*. Amsterdam: John Benjamins. <https://doi.org/10.1075/slsi.29> .
- FORD, Cecilia E., Barbara. A. FOX & Sandra A. THOMPSON. 1996. Practices in the Construction of Turns: The 'TCU' revisited. *Pragmatics* 6(3) : 427-454.
- FOSTER, Pauline, Alan TONKYN & Gillian WIGGLESWORTH. 2000. Measuring spoken language: A unit for all reasons. *Applied Linguistics* 21 : 354–375.

Biography

Valentina Saccone is a postdoctoral researcher in Linguistics at the University of Florence since 2021. Her research primarily focuses on spontaneous communication and the prosody/pragmatics interface, applied to both typical and atypical individuals. Her work encompasses the study of spoken language and its comparison with written text.

Shuai Luo is a doctoral candidate in Linguistics. Her research interests encompass prosody, pragmatics, and corpus linguistics. Her prior publications have delved into the intricacies of information structure in spontaneous spoken language.

Lorenzo Gregori has a PhD in Computer Engineering and a Master in Data Science and Statistical Learning. Since 2011, he works on computational linguistics, and his main fields of interest are semantics, ontologies, and language corpus processing. Nowadays, he is directing his studies towards Large Language Models and the extraction of semantic information from multimodal resources.

LES INDICES MIMIQUE-GESTUELS COMME MARQUEURS DES OPERATIONS ENONCIATIVES

Elena VLADIMIRSKA

Université de Lettonie

jelena.vladimirska@lu.lv

Nous proposons une étude des indices mimique-gestuels en rapport avec la variation sémantique des unités lexicales. Nous soutenons que les variations sémantiques d'une unité lexicale peuvent être associées à une certaine expression gestuelle du locuteur, laquelle, en jonction avec la prosodie et la direction du regard, organise le type particulier de la scène énonciative (Paillard, 2017) propre à cette unité. Ainsi, les gestes et la direction du regard lors de la colocation constituent-ils des traces des opérations cognitives tout comme les éléments lexicaux ou la prosodie. Nous nous référons ainsi au concept de geste mental introduit par A. Culioli qui, en parlant des « traces enfouies dans l'entrelacs du langage », insiste sur le rapport entre notre système sensoriel d'une part, et certaines opérations mentales d'autre part. Plus précisément, à travers notre activité sensorimotrice, on voit apparaître « une mobilisation de nos représentations (donc d'ordre mental), qui est insérée dans des situations existantes (donc : référenciation, au sens de « situer quelque chose », et non de recherche de référent), et toujours accompagnée de régulation – par rapport à ce que l'on est en train de faire, les commentaires qu'on en fait, les controverses avec autrui, etc. » (Culioli 2011 ; 2018 : 62-63).

Notre corpus est constitué à partir de la base de données audiovisuelles © YOUGLISH comportant des situations de communication variées (entretiens, émissions débats, productions des vidéastes, conférences publiques, etc.). Les tracés mélodiques, qui visualisent le mouvement du fondamental, les variations de l'intensité et de la durée, sont obtenus grâce au logiciel de traitement du son Speech Analyzer . Les images sont traitées avec l'utilisation du logiciel d'annotation de fichiers multimédias ELAN . Pour l'étude des indices mimique-gestuels, nous partons des paramètres définis pour la langue des signes française (Bouvet, 1992 : 23) que nous avons adaptés pour notre analyse. On prend en considération 1) la configuration de la main ; 2) l'orientation de la paume ; 3) l'emplacement, c'est-à-dire l'espace où le geste se réalise : dans l'espace du locuteur, dans l'espace partagé des locuteurs, en dehors de l'espace de la colocation ; 4) la trajectoire du geste : sa direction et sa vitesse. Par ailleurs, nous tenons compte de la direction du regard du locuteur et du moment où le regard quitte/revient vers le colocuteur.

À partir des conjonctions régulières des indices prosodiques et mimique-gestuels observées lors de la réalisation des unités lexicales en question, nous distinguons des configurations qui renvoient aux différentes propriétés sémantiques et énonciatives de ces unités. Nos études antérieures portaient sur des marqueurs tels que vraiment (Morel&Vladimirska, 2014), une sorte de, une espèce de, un genre de (Vladimirska&Turlā-Pastare, 2022), n'importe qui/n'importe quoi (Vladimirska, 2024 à par.). Dans la continuation de cette recherche, nous proposons ici une étude du mot esprit en ébauchant par ailleurs quelques pistes pour la typologie des indices mimique-gestuels.

Mots clés : opérations énonciatives, indices mimique-gestuels, sémantique, variation

Références

- BOUVET, Danielle. 1992. « Classification articulatoire des configurations de la main dans la langue des signes française : portée heuristique de cette classification pour la recherche des unités distinctives », *Portée* vol.20, n0 2-3 : 23-32.
- CULIOLI, Antoine. 2018. *Pour une linguistique de l'énonciation*, Tome 4. Tours et détours. Limoges : Lambert Lucas.
- CULIOLI, Antoine. 2011. *Gestes mentaux et réseaux symboliques : à la recherche des traces enfouies dans l'entrelacs du langage* (Conférence retranscrite par R. Camus).
- MOREL, Marie-Annick, et Elena VLADIMIRSKA. 2014. « Intonation and gesture in the segmentation of speech units: The discursive marker vraiment: integration, focalisation, formulation ». *Discourse Segmentation in Romance Languages*, S. Pons Boerderia (dir.), Amsterdam: John Benjamins. 185-218.
- PAILLARD, Denis. 2017. « Scène énonciative et types de marqueurs discursifs ». *Langages*, vol. 207 no. 3 : 17-32.
- VLADIMIRSKA, Elena, et Daina TURLĀ-PASTARE. 2022. « Une sorte de, une espèce de, un genre de : une approche multidimensionnelle des marqueurs dits 'd'approximation' (Sémantique, prosodie, regard, gestes) ». *Information Grammaticale* no 173 : 48 – 60.
- VLADIMIRSKA, Elena (2024 à par.). « Les indices prosodiques et mimique-gestuels comme marqueurs des opérations énonciatives. Exemple des indéfinis n'importe qui et n'importe quoi ». *Actes du Colloque de la TOPÉ*,.Tours.

Notice bio-bibliographique

Elena Vladimirska est docteur en sciences du langage de l'Université Paris 3 Sorbonne Nouvelle et Professeure des universités à l'Université de Lettonie où elle dirige le programme de Master de langues et cultures romanes et le programme de Licence de Philologie Française. Sa recherche porte sur la sémantique, la prosodie et le mimique-gestuel, et est orientée vers le français oral et les études contrastives plurilingues. Elle est l'auteure de l'ouvrage *Exclamation dans le dialogue oral* (Ophrys, 2005), et de nombreux articles et chapitres d'ouvrages sur les marqueurs discursifs, leur sémantique et la réalisation prosodique et mimique-gestuelle dans le discours, parmi lesquels Morel M.-A., Vladimirska E. (2014), *Intonation and gesture in the segmentation of speech units: The discursive marker vraiment: integration, focalisation, formulation*, in *Discourse Segmentation in Romance Languages*, S. Pons Boerderia (dir.), Amsterdam: John Benjamins, pp.185-218 ; Vladimirska E., Turlā-Pastare D. (2022), *Une sorte de, une espèce de, un genre de : une approche multidimensionnelle des marqueurs dits 'd'approximation'*. (Sémantique, prosodie, regard, gestes), *Information Grammaticale* N0 173, pp.48 – 60. Elle est également coéditrice des numéros thématiques des revues (Capin, D., Benninger, C., Vladimirska, E. (eds.), (2020), *De la catégorisation subjective : Procédés et pratiques linguistiques*, *Langue française* N°207 (2/2020) ; Vladimirska E., Romand, D. (eds.), (2022), *Langage et affectivité : perspectives épistémologiques et linguistiques*, *Synergies pays riverains de la Baltique* N°16, et des ouvrages collectifs (Vladimirska, E, Ponchon, T. (eds.), (2016), *Dire l'autre, voir autrui. L'altérité dans la langue et les discours*. Paris, L'Harmattan).