Metaphors we eat by : une analyse du corpus culinaire de la presse française

Joanna Gronkowska
Doctoral School in the Humanities, Université Jagelonne de Cracovie
joanna.gronkowska@doctoral.uj.edu.pl

Introduction

La recherche présentée s'inscrit dans le cadre de l'analyse du discours culinaire français à travers l'étude de métaphores conceptuelles dans la presse contemporaine. En s'inspirant de l'aphorisme célèbre de Brillat-Savarin, « Le destin des nations dépend de la manière dont elles se nourrissent » (1848 : 9), cette étude examine comment la langue et l'identité culinaire française se façonnent mutuellement dans le discours médiatique contemporain.

Corpus et méthodologie

Corpus

Le corpus étudié comprend 413 000 mots provenant d'articles thématiques de trois journaux français majeurs : Le Figaro (197 000 mots), Le Monde (148 000 mots) et Libération (68 000 mots). Bien que le corpus ne soit pas équilibré en termes de distribution entre les journaux, cette asymétrie n'affecte pas l'analyse dans la mesure où l'étude ne vise pas à comparer les orientations politiques ou les tendances des différents journaux.
La collecte des données a été effectuée par *web scraping* et l'analyse a été menée à l'aide de Sketch Engine, un outil d'analyse de corpus en ligne.

Méthodologie

L'approche théorique adoptée s'appuie sur la théorie de la métaphore conceptuelle développée par Lakoff et Johnson (1980). Cette théorie postule que les métaphores ne sont pas simplement des ornements linguistiques mais constituent des mécanismes fondamentaux de l'esprit qui permettent de comprendre des domaines plus ou moins abstraits en termes de domaines plus concrets et familiers. La métaphore structure ainsi notre compréhension de l'expérience et influence nos perceptions et actions de manière souvent inconsciente.

La méthodologie employée suit les principes de la linguistique de corpus, particulièrement l'approche corpus-driven Deignan (2005 : 88). Cette approche se distingue de l'approche corpus-based en ce qu'elle ne part pas d'hypothèses préétablies, mais laisse émerger les observations directement des données du corpus. Pour l'identification et l'extraction des données métaphoriques, l'étude applique la méthode proposée par Stefanowitsch (2007 : 2). Parmi les cinq méthodes qu'il énumère pour identifier les métaphores dans les corpus : (1) recherche manuelle, (2) recherche de vocabulaire du domaine source, (3) recherche de vocabulaire du domaine cible, (4) recherche de phrases contenant des éléments lexicaux des deux domaines, et (5) recherche basée sur les « marqueurs de métaphore ». Cette recherche utilise surtout la deuxième approche, centrée sur le vocabulaire du domaine source. Cinq mots-clés ont été sélectionnés : produit, terroir, recette, plat, et repas. Ce choix reflète symboliquement le parcours du produit vers la table, tout en proposant des concepts de nature différente. Le terroir représente un concept complexe ancré dans la géographie et la culture, la recette constitue une composition des produits, le plat est la conséquence directe d'une recette, et le repas appartient davantage à l'identité sociale. Comme le souligne Stefanowitsch (2007 : 3), le choix des éléments peut être basé sur des décisions a priori, limité aux contextes prometteurs ou simplement pertinents pour les questions de recherche.
Nous formulons donc une question de recherche principale : de quelle manière les métaphores présentes dans le discours culinaire participent-elles à construire et à exprimer la cuisine comme un symbole majeur de l'identité nationale française ?

Résultats

L'analyse préliminaire révèle l'émergence de plusieurs domaines métaphoriques associés au vocabulaire culinaire. Les unités linguistiques étudiées sont notamment associées aux domaines de la guerre, de la musique, du langage, du théâtre et du récit.

La cuisine est représentée comme un champ de bataille où on « défend le produit », où un « chef terroiriste » incarne un militant du patrimoine, et où l'on mobilise « une armée des plats » pour relever les défis culinaires. Ensuite, la cuisine est conçue comme un langage vivant : on peut « conjuguer le terroir », « décliner une recette » ou encore « dicter un repas ». Le vocabulaire culinaire emprunte aussi à la narration : on « raconte », « présente » ou « propose un produit ». Le produit ou le repas deviennent ainsi un support de communication et de mémoire. La cuisine est comme une œuvre musicale où « clore le repas sur une note » signifie

finir le repas de façon harmonieuse, mettant en avant le côté esthétique et agréable de l'expérience culinaire, tout comme une symphonie se termine par une belle mélodie. Certaines expressions évoquent la dramaturgie avec des termes comme « comme en scène », « mise en scène culinaire », ou « la gastronomie théâtrale », soulignant le caractère performatif des pratiques culinaires.

L'interprétation des résultats s'inscrit dans l'approche pluridisciplinaire des Food Studies, telle que conceptualisée par Albala (2013) : l'étude de l'alimentation nécessite la mobilisation de plusieurs disciplines pour saisir la complexité des phénomènes culinaires dans leurs dimensions sociales, culturelles, historiques et linguistiques.

Cette analyse du corpus culinaire de la presse française révèle la richesse et la complexité des métaphores qui structurent le discours contemporain sur l'alimentation. Elle contribue à une meilleure compréhension des mécanismes par lesquels la langue façonne notre perception. Les premiers résultats suggèrent que le discours culinaire français contemporain mobilise un réseau complexe de métaphores qui dépassent le simple domaine alimentaire pour toucher aux dimensions culturelles, artistiques et sociales de l'expérience culinaire.

Références bibliographiques

Albala, K. (2013). *Routledge international handbook of food studies* (1st ed.). Routledge. https://doi.org/10.4324/9780203819227

Barthes, R. (2012). Toward a psychosociology of contemporary food consumption. In *Food and culture* (3rd ed.). Routledge.

Brillat-Savarin (1848). *Physiologie du gout.*

Deignan, A. (2005). *Metaphor and corpus linguistics*. John Benjamins Publishing Company.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Stefanowitsch, A., & Gries, S. (2007). *Corpus-based approaches to metaphor and metonymy*. De Gruyter Mouton. https://doi.org/10.1515/9783110199895

Analyser les échanges médecins-patients autour des termes médicaux et procéduraux : un projet en question(s)

Florine Voisard , Aurélie Picton

Département de traitement informatique multilingue [Genève] (TIM)

Ce projet a pour but d'étudier les corpus oraux en terminologie et langues de spécialité afin de décrire les termes et leur circulation en contexte oral, participant ainsi à la réflexion initiée sur la construction d'une terminologie dite interactionnelle (Giraudier, 2024). Dans le cadre de situations de prise en charge d'urgence, il s'agira d'identifier les termes médicaux et connaissances spécialisées partagés entre personnel soignant et personne soignée et d'en analyser les usages. Porté sur le contexte suisse, encore peu exploré, le projet vise à mettre en évidence les enjeux terminologiques de la communication médicale lors d'urgences, à comprendre les conditions dans lesquelles certains choix sont faits et perçus, et à identifier les pratiques favorisant une communication claire, bienveillante et efficace. Les résultats contribueront également à alimenter la réflexion sur les corpus spécialisés oraux, puisque le projet s'appuie sur un corpus de données orales et transcrites.

Introduction

Cette communication, sous forme de poster, vise à présenter un projet de thèse de doctorat qui débute au sein de notre groupe de recherche. Ce projet propose d'aborder la question des interactions entre médecins et patients, en situation d'urgences préhospitalières, à travers un double point de vue : interactionnel et terminologique. Ceci ouvre différentes questions théoriques et méthodologiques, qui contribuent à alimenter les réflexions sur l'analyse des interactions en corpus spécialisés, et notamment l'idée d'une "terminologie interactionnelle" (Giraudier 2024, Giraudier & Riou 2024).

Dans le cas des urgences médicales, les enjeux autour de la communication avec les patients sont importants (par ex. Riou 2024). Toutefois, l'asymétrie de savoirs entre le spécialiste du domaine (personnel médical) et le non-spécialiste (personne prise en charge) peut venir troubler la clarté de la communication (voir par ex. Fillietaz 2006 ; Vaajala et al. 2013 ; Schmucki 2018). Différents points de vue participent à documenter ce type d'échanges complexes en linguistique (par ex. Skelton 2008 ; Rollet 2015 ; Singy & Merminod 2021 pour des points de vue sociolinguistiques, Delavigne 2007, 2008 pour un regard socioterminologique ou encore Drew et al. 2001 ; Heritage & Maynard 2006 ; Giraudier et Riou 2024, en linguistique interactionnelle). Cependant, bien que parfois très proches et complémentaires, force est de constater que ces différents points de vue ne "dialoguent" pas toujours réellement.

Récemment, Giraudier (en cours) a ainsi mis en évidence la complémentarité entre les perspectives interactionnelles et une vision orientée *terminologie*, considérant ainsi l'interaction comme un discours spécialisé, où circulent les connaissances et termes du domaine médical entre groupes de locuteurs : le corps médical et les patients. La linguistique interactionnelle a considérablement documenté les échanges en situation d'urgence, par exemple les appels téléphoniques aux services de secours (Riou 2024) ou, plus ponctuellement, les interactions en contexte préhospitalier (Auvinen & Palukka 2012 ; Deppermann 2014 ; Marzuki et al. 2020). Toutefois, elle s'est plus discrètement intéressée à la circulation précise des termes spécialisés dans ces échanges. De son côté, la terminologie a développé des outils solides pour décrire les mécanismes de formation, de variation et de circulation des termes (Delavigne & Gaudin 2022 ; Humbert-Droz & Picton 2024), mais en se fondant le plus souvent sur des corpus écrits, éloignés des usages réels en interaction. Or, dans les contextes d'urgence, où la

terminologie participe directement à la coordination et à la prise en charge des patients (Lerner et al. 2000 ; Hayes et al. 2017), croiser ces deux approches ouvre la possibilité d'analyser finement les usages oraux des discours spécialisés et de mieux comprendre la circulation des savoirs en situation.

Dans ce contexte, notre projet vise à proposer un cadre méthodologique d'observation et une description de la circulation des termes et concepts médicaux dans des situations de prise en charge de patients et à contribuer ainsi à la réflexion initiée sur la construction d'une terminologie dite interactionnelle. Nous nous intéressons en particulier au contexte suisse, marqué par une grande expertise médicale dans le transport d'urgences (en particulier héliportées), potentiellement en situation linguistique plurilingue (par ex. Perera et al. 2025), mais encore très peu observé.

La mise en place de notre recherche, dans ce contexte complexe, pose différentes questions d'ordre théorique et méthodologique que notre projet commence à articuler, et que nous souhaitons partager ici. Dans cette communication, nous proposons de décrire les principaux questionnements méthodologiques soulevés quant à la constitution des données d'analyse. Puis, nous présentons l'état actuel de notre collecte. Enfin, nous partageons les principaux axes de réflexion et de discussion ouverts, que notre recherche contribuera, nous l'espérons, à alimenter dans les années à venir.

Corpus et méthodologie

Corpus : spécifications et défis

La collecte des données dans ce type de recherche est difficile pour différentes raisons (Descamps 2022 ; Riou 2024), dont la difficulté d'accès à des données réelles auprès de partenaires, la difficulté technique à enregistrer, la lourdeur d'annotation/transcription de ces données, ainsi que le processus essentiel de validation éthique de ce type de protocole qui touche à l'expérience médicale. À ces difficultés pratiques s'ajoutent différents questionnements méthodologiques quant à la nature des données à privilégier pour travailler sur des interactions. En effet, en linguistique interactionnelle, un corpus d'un nombre restreint d'interactions peut suffire pour une analyse, là où une approche terminologique nécessite généralement des corpus plus conséquents (Bowker & Pearson 2002).

Dans notre projet, nous faisons appel à différents cadres méthodologiques pour soutenir notre analyse : 1. puisque nous nous intéressons à la circulation des savoirs et des termes, notre approche peut reposer sur des propositions de la terminologie textuelle (Condamines & Picton 2022) et de la socioterminologie (Delavigne & Gaudin 2022) et privilégie des corpus, écrits ou transcrits, comparables spécialisés, ainsi que des entretiens ; 2. du fait de son approche orale, notre étude s'ancre dans l'étude des interactions et peut bénéficier des propositions méthodologiques de la linguistique interactionnelle sur l'observation de données orales, leurs spécificités et leurs transcriptions.

Méthodologie de constitution du corpus de transcriptions : état des lieux

Pour construire cette articulation, et sur la base des travaux existants à ce stade, nous ciblons quatre types de données (Table 1). Nous collaborons actuellement avec un service d'urgences héliportées et avons commencé à récolter des données d'observation. En parallèle, nous continuons à échanger avec d'autres partenaires potentiels.

Type de données	**Apports attendus de ces données**	**Approches pour la constitution**
corpus d'interactions orales transcrites entre personnel médical et personnes prises en charge Objectif : 100-150 échanges d'interventions (une intervention = 10 min.)	Rendre compte des interactions sur le terrain, dans leur environnement naturel	Enregistrement des données par le partenaire, accord de partage des transcriptions Transcription des données (*Conventions de transcription – CORINTE*. (s. d.) et Hagemann, J., & Henle, J. (2021))

Langues : allemand, français et anglais possibles Statut : En cours d'enregistrement Validation éthique reçue.		Variables prises en compte : gravité du cas, tranche d'âge, langue d'intervention, langue de culture du médecin et de la personne prise en charge, spécialisation du médecin, années d'expérience dans les secours
corpus de rapports d'interventions (lorsque disponibles) liés à ces prises en charge Statut : En cours d'autorisation.	Compléter le corpus d'interactions orales avec une description factuelle du cas et de la terminologie	Autorisation de partage du canton, anonymisé Conversion PDF
données d'observation de terrain (carnet, photos, prise de notes) Statut : Réalisées.	Affiner la compréhension du terrain de la collecte des données, l'organisation des équipes sur le terrain, la logistique spatiale (équipements et personnes)	Observation en direct, participation à plusieurs interventions
données d'entretien avec le personnel médical Statut : À venir en aval de la récolte de données transcrites.	Apporter le point de vue du personnel médical sur les prises en charge avec les patients, leurs impressions, ressentis et expériences au niveau de la communication avec les patients	Focus group
données de questionnaire à destination des personnes prises en charge Statut : À venir en aval de la récolte de données transcrites.	Apporter le point de vue des patients sur les prises en charge, leurs impressions, ressentis au niveau de la communication avec les équipes médicales	Questionnaire en ligne

table 1. : Type de données, en cours de construction

Discussions et principaux résultats attendus

À travers ce poster, nous souhaitons partager les premiers axes de réflexion mis en évidence dans ce projet de thèse, qui s'inscrit dans la lignée des approches en corpus pour l'analyse des interactions.

Dans un premier temps, l'axe principal de discussion repose sur la proposition d'une description systématique de l'usage de la terminologie dans les échanges médecin/patient en Suisse. Notre objectif est donc d'identifier les termes et connaissances spécialisées partagés et d'analyser leurs usages dans ce type de cadre communicationnel. Plus précisément, nous cherchons à mettre au jour les fonctionnements linguistiques en jeu, ainsi que les stratégies de production et de reformulation des termes et expression des gestes médicaux dans ce cadre d'interaction marqué par une asymétrie de savoirs, dans un contexte d'urgence, de stress, ou de nécessité pédagogique. Nous proposons de nous inspirer notamment des travaux sur les marqueurs définitoires (Meyer 2001 ; Vergely et al. 2009 ; Traverso 2022), les stratégies de marquage des territoires épistémiques (par ex. Gradoux 2021 ; Fillietaz & Zuppiger 2023), l'utilisation du métalangage (par ex. Lecolle et al. 2018 ; Delavigne 2020) ou, plus généralement, l'analyse de contextes distributionnels (par ex. Condamines & Picton 2022).

Ces descriptions visent à enrichir mutuellement les perspectives interactionnelles et les travaux en langues de spécialité/terminologie, prolongeant ainsi la réflexion sur une "terminologie interactionnelle" (Giraudier, 2024). Les résultats attendus de cette étude visent à répondre à plusieurs questionnements théoriques et méthodologiques, dont la place des corpus oraux d'interactions pour une étude située de la terminologie. Cet enjeu va de pair avec le questionnement de plusieurs enjeux tels que la nature et la taille des données, leur complémentarité, auquel s'ajoute la question de l'équilibre entre les besoins de la recherche et la réalité des données de terrain que l'on peut récolter. Notre cadre de travail amène également à s'interroger sur la dimension multilingue, dont on sait la présence et le rôle certain (par ex. Grin 2010, Perera et al. 2025), mais qui reste difficile à anticiper précisément dans les échanges

récoltables. Il sera donc nécessaire d'évaluer et questionner la mesure dans laquelle le rôle du multilinguisme peut être observé dans cette réalité des données.

Enfin, d'un point de vue appliqué, ce projet d'analyse fine des interactions contribuera à identifier les pratiques favorisant une communication bienveillante et efficace, tout en repérant les obstacles ou les atouts terminologiques éventuels. Ces résultats pourront nourrir des pistes concrètes d'amélioration en matière de formation, de médiation linguistique et de communication en santé.

Références bibliographiques

Auvinen, P., & Palukka, H. (2012). Organization of work in interaction between the paramedics and the patient. *Work 41, (S1)*, 42-48.

Bowker, L., & Pearson, J. (2002). *Working with specialized language*. Routledge.

Condamines, A., & Picton, A. (2022). Textual Terminology : Origins, principles and new challenges. In P. Faber & M.-C. L'Homme (Éds.), *Terminology and Lexicography Research and Practice* (Vol. 23, p. 219-236). John Benjamins Publishing Company.

Conventions de transcription – *CORINTE*. (s. d.). Consulté le 28 avril 2025, à l'adresse https://icar.cnrs.fr/corinte/conventions-de-transcription/

Delavigne, V. (2007). Les mots des patients atteints de cancer. Socioterminologie et pratique professionnelle en santé. *Les langues de spécialité en question : perspectives d'étude et applications. 12ème journée scientifique de la CRL*, Dardo de Vecchi; Claire Martinot, Nov 2007, Paris, France, 18-32. https://hal.science/hal-00920651v1

Delavigne, V. (2008). Des termes des experts aux mots des patients : Un discours à construire. *Analyse de discours et demande sociale : enjeux théoriques et méthodologiques*, Nov 2008, Paris, France. https://hal.science/hal-00920760v1

Delavigne, V. (2020). De l'(In)constance du métalinguistique dans un corpus de vulgarisation médicale. *Corela. Cognition, représentation, langage*, *HS-31*. https://doi.org/10.4000/corela.11031

Delavigne, V., & Gaudin, F. (2022). Founding principles of Socioterminology. In P. Faber & M.-C. L'Homme (Éds.), *Terminology and Lexicography Research and Practice* (Vol. 23, p. 177-196). John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.23.08del

Deppermann, A. (2014). Multimodal participation in simultaneous joint projects Interpersonal and intrapersonal coordination in paramedic emergency drills. In *Multiactivity in Social Interaction* (John Benjamins, p. 247-281). https://doi.org/10.1075/z.187

Descamps, F. (2022). Éditorial. *Bulletin de l'AFAS. Sonorités*, *48*, 4-5. https://doi.org/10.4000/afas.7314

Drew, P., Chatwin, J., & Collins, S. (2001). Conversation analysis : A method for research into interactions between patients and health-care professionals. *Health Expectations, 4*(1), 58-70. https://doi.org/10.1046/j.1369-6513.2001.00125.x

Filliettaz, L. (2006). Asymétrie et prises de rôles. Le cas des réclamations dans les interactions de service. In *Les interactions asymétriques* (Québec : Editions Nota bene, p. 89-112). https://archive-ouverte.unige.ch/unige:37651

Filliettaz, L., & Zuppiger, J. (2023). Négocier les rapports épistémiques dans les entretiens de bilan en crèche : Le cas des actions de conseil entre les parents et les professionnels de l'éducation de l'enfance. *Éducation et socialisation*, *67*. https://doi.org/10.4000/edso.23236

Giraudier, E. (en cours). *Linguistique et médecine d'urgence : analyse terminologique et interactionnelle des appels d'urgence au SAMU 69 pour traumatismes graves*. Thèse de doctorat en linguistique anglaise, Université Lyon 2.

Giraudier, E. (2024). *La « terminologie interactionnelle », un cadre innovant pour l'étude des termes dans les appels d'urgence téléphonique au SAMU*. 13e journées scientifiques du réseau LTT - Lexicologie, Terminologie, Traduction, Paris.

Giraudier, E., & Riou, M. (2024). « Petit malaise » ou « très gros souci » ? Atténuation et intensification dans les descriptions profanes d'urgences vitales au téléphone. *SHS Web Conf. Volume 191, 2024*. Congrès Mondial de Linguistique Française 2024, Lausanne. https://doi.org/10.1051/shsconf/202419101023

Gradoux, X. (2015). Expertises partagées en médecine générale : Orientations épistémiques vers le symptôme et le diagnostic. *Cahiers du Centre de Linguistique et des Sciences du Langage*, *42*, 1131. https://doi.org/10.26034/la.cdclsl.2015.635

Grin, F. (2010). L'aménagement linguistique en Suisse. *Télescope, 16(3),* 55–74.

Hagemann, J., & Henle, J. (2021). *Transkribieren nach GAT 2 (Minimal- und Basistranskript) – Schritt für Schritt. Version actualisée*. Pädagogische Hochschule Freiburg. https://www.ph-freiburg.de/fileadmin/shares/Projekte/Quasus/Dateien/Transkribieren_nach_GAT_2_-_Schritt_fuer_Schritt__Aktualisierte_Version_.pdf

Hayes, E., Dua, R., Yeung, E., & Fan, K. (2017, décembre 1). Patient understanding of commonly used oral medicine terminology. *british dental journal*, 842-845.

Heritage, J., & Maynard, D. W. (2006). Introduction : Analyzing interaction between doctors and patients in primary care encounters. In J. Heritage & D. W. Maynard (Éds.), *Communication in Medical Care* (1re éd., p. 121). Cambridge University Press. https://doi.org/10.1017/CBO9780511607172.003

Humbert-Droz, J., & Picton, A. (2024). *Révéler l'expertise partagée par les patientes atteintes de diabète et d'endométriose : Une analyse exploratoire de forums médicaux*. Actes des 17es Journées internationales d'Analyse statistique des Données Textuelles/International Conference on the Statistical Analysis of Textual Data. http://archive-ouverte.unige.ch:/unige:180056

Lecolle, M., Veniard, M., & Guérin, O. (2018). Pour une sémantique discursive : Propositions et illustrations: *Langages*, *N° 210*(2), 3554. https://doi.org/10.3917/lang.210.0035

Lerner, E. B., Jehle, D. V. K., Janicke, D. M., & Moscati, R. M. (2000). Medical communication: Do our patients understand? *The American Journal of Emergency Medicine*, *18*(7), 764-766.

Marzuki, E., Rohde, H., Cummins, C., Branigan, H., Clegg, G., Crawford, A., & MacInnes, L. (2020). Closed-loop communication during out-of-hospital resuscitation. *Communication & Medicine*, *16*(1), 54-66. https://doi.org/10.1558/cam.37034

Meyer, I. (2001). Extracting knowledge-rich contexts for terminography : A conceptual and methodological framework. In D. Bourigault, C. Jacquemin, & M.-C. L'Homme (Éds.), *Natural Language Processing* (Vol. 2, p. 279-302). John Benjamins Publishing Company.

Perera, N., Riou, M., Birnie, T., Whiteside, A., Ball, S., & Finn, J. (2025). Language barriers in emergency ambulance calls for cardiac arrest: Cases of missing vital information. *Social Science & Medicine, Volume 365*. https://doi.org/10.1016/j.socscimed.2024.117623.

Riou, M. (2024). Communication in Prehospital and Emergency Care: A State-of-the-Art Literature Review of Conversation-Analytic Research. *Research on Language and Social Interaction*, *57*(1), 55-72. https://doi.org/10.1080/08351813.2024.2305044

Rollet, N. (2015). Conversations dans l'urgence. *Multitudes, 60(3)*, 87–93.

Schmucki, A. (2018). *Placeboreaktionen und Noceboeffekte im Rettungsdienst oder Tue Gutes und sprich darüber ! - Anregungen zu einer positiv suggestiven Kommunikation in der Präklinik* [Diplomarbeit]. Höheren Fachschule für Rettungsberufe HFRB.

Singy, P., & Merminod, G. (2021). *La communication en milieu médical : Un labyrinthe*. Presses polytechniques et universitaires romandes.

Skelton, J. (2008). *Language and Clinical Communication*. Oxford: Radcliffe Publishing.

Traverso, V. (2022). Les définitions et leurs enjeux dans des consultations médicales avec des demandeurs d'asile. In V. Montagne (Éd.), *Stratégies de la définition*. ⟨halshs-03902129⟩

Vaajala, T., Arminen, I., & De Rycker, A. (2013). Misalignments in Finnish emergency call openings : Legitimacy, asymmetries and multi-tasking as interactional contests. In A. De Rycker & Z. Mohd Don (Éds.), *Discourse Approaches to Politics, Society and Culture* (Vol. 52, p. 131157). John Benjamins Publishing Company. https://doi.org/10.1075/dapsac.52.04vaa

Vergely, P., Condamines, A., Fabre, C., Josselin-Leray, A., Rebeyrolle, J., & Tanguy, L. (2009). Analyse linguistique des interactions patient/médecin. *Actes éducatifs et de soins, 2009, Nice, France (publication numérique)*.

CORLI CORPUCIT

Un outil web pour créer des citations pérennes d'extraits de corpus ou de textes et les insérer dans vos articles scientifiques et livres au format PDF

Consortium CORLI
Driss Sadoun (PostLab & ERTIM/INALCO)
Christophe Parisse (CORLI & Modyco - CNRS & Université Paris Nanterre)

Introduction

Le développement de la publication électronique permet d'enrichir très largement la citation d'extrait de langage ou de corpus. On peut utiliser de multiples formats et présentations, mettre du son, de l'image, de la vidéo. Mais ces présentations sont éphémères à la différence du support papier.

CORPUCIT est un outil qui permet de créer des citations d'extraits de corpus ou de texte, pérennisées sur NAKALA, insérées dans vos articles ou ouvrages, et qui font référence aux corpus ou sources originales. Les citations étant électroniques, il n'y a pas de limite de format.

CORPUCIT permet de favoriser la science ouverte, la reproduction des travaux scientifiques par la pérennité des données décrites. Il permet aussi de faire connaître l'usage des corpus sur Internet.

Projet CITATION : CORPUCIT
Faciliter la citation et la réutilisation d'extraits de corpus langagiers

Christophe Parisse (INSERM, MoDyCo - Université Paris Nanterre)
Driss Sadoun (PostLab, ERTIM - INALCO)

Contexte et problématique

Dans les publications scientifiques en linguistique et sciences du langage, les extraits de corpus jouent un rôle central :

- Ils illustrent concrètement les phénomènes linguistiques étudiés
- Ils ancrent les analyses théoriques dans des données réelles
- Ils permettent la vérification et la discussion des interprétations
- Ils constituent des preuves empiriques à l'appui des démonstrations

Problèmes actuels

- Absence de lien direct entre la publication et les données
- Extraits décontextualisés, difficiles à retrouver dans le document source
- Difficulté de partage et diffusion
- Absence de crédit des personnes à l'origine des extraits
- Manque de traçabilité et de pérennité
- Non reconnaissance comme matériels scientifiques à part entière

Ambition

Mettre à disposition une plateforme permettant de lier finement écrits scientifiques et extraits de données de langue (écrits, sons, vidéos, images), présentées dans leur contexte, facilitant la réflexion scientifique et la réutilisation des données.

La plateforme CORPUCIT vise à accompagner les chercheurs pour :

- Simplifier la création d'extraits (exemples et illustrations) pouvant être insérés dans des écrits scientifiques ou pédagogiques.
- Archiver les extraits de manière pérenne.
- Générer des citations pouvant être insérées dans des écrits scientifiques.
- Visualiser, contextualiser et manipuler des extraits de corpus.
- Valoriser la création et le partage d'extraits comme activité scientifique
- Favoriser la science ouverte et la reproductibilité des travaux

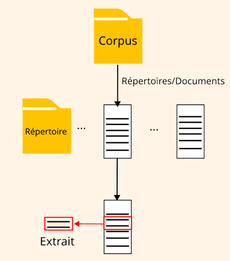
Extraits de corpus : définition et exemples

Définition d'un extrait dans le contexte de CORPUCIT

Passage ou portion d'un document (écrit, audio, vidéo ...) faisant partie d'un corpus langagier.

On distingue trois objets :

- Corpus : peut contenir des documents et des répertoires.
- Document : fichier texte, image, audio ou vidéo.
- Extrait : tout ou partie d'un document.



Méthode de préservation et de diffusion des extraits

- L'interface avec NAKALA permet de bénéficier de ses mécanismes d'archivage.
- Les extraits existent indépendamment de CORPUCIT.
- CORPUCIT est donc une interface de création et d'accès à une base de données d'extraits pérennes.

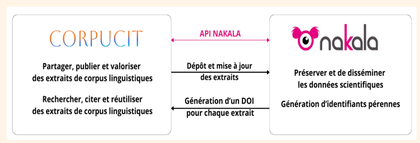


Figure: Interactions entre CORPUCIT et NAKALA

Conclusion

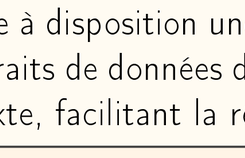
Les extraits de corpus langagiers permettent d'illustrer concrètement des phénomènes linguistiques, pour soutenir ou expliquer un propos scientifique. Un extrait est donc un matériel scientifique.

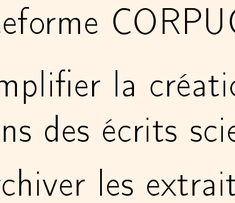
Nous proposons une plateforme de gestion d'extraits de corpus avec pour objet :

- Simplifier la création, visualisation, manipulation et réutilisation d'extraits.
- Offrir au lecteur un accès direct à l'extrait.
- Contextualiser l'extrait dans le document ou le corpus original.
- Favoriser le crédit des personnes ayant créé ou identifié un extrait.
- Respecter les principes FAIR et participer la diffusion scientifique.

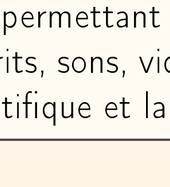
Flux de gestion des extraits dans la plateforme CORPUCIT : de la création à la citation

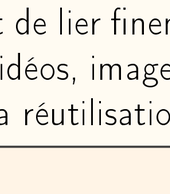
Explorer et rechercher

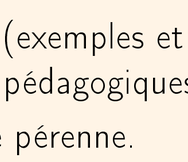


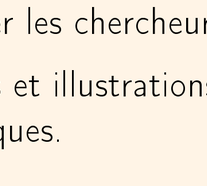


Créer et éditer

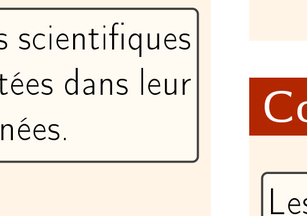




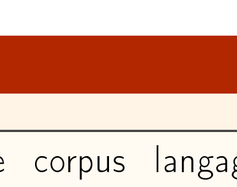


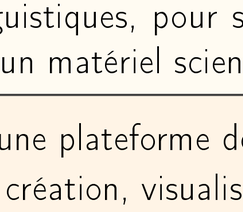


Visualiser



Archiver dans NAKALA





Illustrer et citer

15 Nous voyons à travers cet exemple que la méthodologie de gestion du bruit lors de la phase de mise en forme influe sur celle de la phase de traitement. Les deux exemples ci-dessous illustrent ce phénomène. Dans l'exemple 4, le choix a également été fait de ne pas surinterpréter la production enfantine. Cependant, on aurait pu considérer que la forme [e] produite par l'enfant devait être interprétée comme un filler prélexical, c'est-à-dire dans ce contexte une forme précoce de pronom personnel. La forme verbale aurait alors sans doute été annotée comme une forme verbale du présent de l'indicatif, comme dans l'exemple 5.

Exemple 4 : Bruit dans l'annotation morphosyntaxique : Impératif (Liégeois 2024d)

*MOT: regarde .
%mor: v|regarder-IMP&2s .

*MOT: qu'est-ce qu'il mange le bébé ?
%mor: pro:int|quelle=int|c'est-ce pro:rel|que&pro:sub|il v|manger-PRES&13s det:art|le&dug n|bébé&fin ?

[...]

*CHI: e@g mange quoi ?
%mor: phon|e v|manger-IMP&2s pro:int|quoi ?

*MOT: un petit suisse !
%mor: det:art|un&ind&sg adj|petit&fem n|suisse !

Exemple 5 : Annotation morphosyntaxique automatique correcte (Liégeois 2024e)

*CHI: on mange .
%mor: pro:sub|on v|manger-PRES&13s .

Liégeois L. (2024d). *Bruit dans l'annotation morphosyntaxique : Impératif* [transcription_video]. CORPUCIT (Huma-Num - CORLI). [https://doi.org/10.34847/nkl.b43evucj]
DOI : 10.34847/nkl.b43evucj

Liégeois L. (2024e). *Annotation morphosyntaxique automatique correcte* [transcription_video]. CORPUCIT (Huma-Num - CORLI). [https://doi.org/10.34847/nkl.2d3fcf64]
DOI : 10.34847/nkl.2d3fcf64

Réoutillage de concordances pour le filtrage et l'échantillonage

Achille Falaise
Laboratoire de Linguistique Formelle (LLF - UMR7110)
Centre National de la Recherche Scientifique, Université Paris Cité

De nombreux outils permettent aujourd'hui d'effectuer des recherches dans des corpus, et de produire des concordances KWIC. On peut par exemple citer AntConc, SketchEngine, et TXM. Toutefois, ces concordances révèlent souvent des problèmes. Il peut par exemple y avoir des erreurs dans le corpus (tokenisation, analyse, inclusions de para-texte, etc.), ou bien les annotations ou le langage de requête ne permettent pas d'extraire précisément ce que l'on désire. Cela entraîne du bruit dans les résultats, qu'il faut bien à un moment mesurer et filtrer. Cela se fait souvent avec un tableur, mais cet outil est peu adapté au traitement de données textuelles.

L'outil présenté dans ce papier tente de répondre à ce besoin. Il s'agit d'un outil en ligne, qui permet de téléverser des concordances (fichiers CSV), et de les retravailler à l'aide de quelques fonctionnalités :

- **Mélange et échantillonnage**. Il est possible de mélanger les concordances avant d'effectuer un échantillonnage. Le mélange est déterministe (pseudo-aléatoire), pour permettre sa reproductibilité. On peut ensuite décider de conserver les n premières lignes pour une analyse approfondie semi-automatique, à l'aide des fonctionnalités suivantes.
- **Analyse des concordances** en POS et lemmes, actuellement avec TreeTagger. L'analyse tient compte de la structuration en trois colonnes des concordances KWIC (fusion → analyse → séparation).
- **Filtrage manuel**. Une case à cocher permet de désactiver individuellement chaque ligne.
- **Filtrage semi-automatique**. Le filtrage peut s'effectuer sur un mot clé, une expression régulière, ou bien, le cas échéant, exploiter l'analyse effectuée par TreeTagger (Schmid, 1994). Dans ce cas, le filtrage peut se faire à l'aide d'un assistant, afin de pouvoir facilement tenir compte des annotations linguistiques.

L'outil permet enfin de télécharger les concordances. Aucune ligne n'est supprimée, mais une colonne est ajoutée, indiquant si la ligne a été filtrée, et le cas échéant par quelle fonctionnalité.

Références bibliographiques

Anthony, L. (2024). AntConc (Version 4.3.1). Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/AntConc

Heiden, S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In 24th Pacific Asia Conference on Language, Information and Computation (p.10). Sendai, Japan. [https://shs.hal.science/halshs-00549764]

Kilgarriff, A.; Rychlý, P.; Smrž, P.; Tugwell, D. (2004). The Sketch Engine.Proceedings of the 11th EURALEX International Congress (p. 105-116).http://www.sketchengine.eu/

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

AVAA Toolkit : une boîte à outils pour accompagner l'analyse des interactions à partir de corpus multimodaux

Clotilde George [1]
[1] Laboratoire ATILF, Université de Lorraine
clotilde.george@univ-lorraine.fr

Introduction

Le logiciel AVAA Toolkit (Audio and Video Annotations Analysis Toolkit) offre de nombreuses fonctionnalités pour l'analyse des interactions à partir de corpus de données audiovisuelles annotées. Conçu initialement comme un logiciel compagnon du logiciel ELAN, son usage est particulièrement adapté à la gestion et à l'exploration de corpus construits avec cet annotateur. AVAA Toolkit peut être mobilisé à différentes étapes du processus de recherche : traitement (transcription automatique, anonymisation de la bande son), fouille (requête des annotations par couche d'annotation, chaîne de caractères ou expression régulière), visualisation de données (collections de données audiovisuelles et leurs annotations sous formes de liste, tableau, timelines, graphiques). Son usage peut également intervenir lors du processus d'annotation lui-même avec une procédure d'intercodage permettant la construction collaborative des items de codage par la comparaison automatique de choix de codage et la production d'un indice d'accord interannotateurs.

Animatrice

Clotilde George (jeune docteure en sciences du langage, associée à l'Université de Lorraine et à l'ATILF).

Sujet

Démonstration du logiciel AVAA Toolkit, adapté à l'analyse de corpus audiovisuels annotés (notamment avec ELAN). Tour de l'interface et des principales fonctionnalités du logiciel.

Ressources (en anglais)

Logiciel : www.avaa-toolkit.org
Présentation rapide : https://avaa-toolkit.org/features/
Documentation : https://avaa-toolkit.org/documentation

Modalités

Durée : 2h

Pas de prérequis.

Est susceptible d'intéresser toute personne impliquée dans une recherche mobilisant un corpus oral ou multimodal a fortiori.

Comment créer, explorer et analyser un corpus de publications scientifiques avec Istex ?

Mathilde Huguin [1]
[1] Institut de l'Information Scientifique et Technique CNRS (UAR76)
mathilde.huguin@inist.fr

L'analyse de corpus de publications scientifiques soulève de nombreux enjeux dont les retombées dépassent la linguistique fondamentale pour toucher aux domaines didactiques, sociaux et politiques. Les travaux dans ce champ de recherche permettent, par exemple, de décrire finement des structures rhétoriques pour aider à la rédaction (Pho, 2008), de construire des lexiques spécialisés à des fins didactiques (Hatier & Tutin, 2019) ou encore d'améliorer la traduction automatique des publications pour rédiger et lire dans sa langue maternelle (Bénard *et al.*, 2023). Ces analyses mobilisent souvent des connaissances avancées en TAL voire sont directement menées par des spécialistes du domaine. L'objectif de cette démonstration est de montrer comment les outils de l'infrastructure Istex [1] (*Initiative d'excellence en information scientifique et technique*) permettent de simplifier la création et l'analyse d'un corpus de publications sans faire appel à des compétences spécialisées en informatique.

Istex est une infrastructure adossée à l'Inist-CNRS poursuivant deux objectifs : **(i)** rendre disponibles des publications scientifiques multilingues et multidisciplinaires via une API, i.e. construire un *réservoir à corpus* selon Habert (2000) ou une *archive* au sens de Rastier (2004) ; **(ii)** faciliter la constitution, l'exploration et l'analyse de corpus construits à partir de ces documents. Aujourd'hui, Istex regroupe plus de 30M d'articles et d'ouvrages en texte intégral [2]. Les données Istex ont fait l'objet de plusieurs traitements visant leur enrichissement (e.g. détection d'entités nommées, catégorisation en domaines scientifiques, Cuxac & Collignon, 2017 ; Cuxac *et al.*, 2017) et leur standardisation (e.g. production de format TEI ou de texte nettoyé). Les enrichissements facilitent la recherche de documents. La standardisation minimise, quant à elle, les pré-traitements nécessaires pour l'analyse du corpus.

Au sein de l'infrastructure, deux outils offrent respectivement de créer et d'explorer un corpus.

- La constitution d'un corpus à partir de l'API est simplifiée par l'application Istex Search [3]. Elle permet de construire une requête et de l'affiner en supprimant le bruit grâce à des filtres, des indicateurs et un accès aux textes intégraux. Istex Search permet également de télécharger massivement les documents au format souhaité. En plus de ce choix personnalisable, quatre formats sont paramétrés pour l'import des données dans des outils d'analyse : CorTexT (Breucker *et al.*, 2016), GarganText (Delanoé & Chavalarias, 2023), Lodex (Gregorio *et al.*, 2019) et NooJ (Silberztein, 2016).
- Lodex [4] (*Linked open data experiment*) est un outil open source permettant l'exploration et l'analyse d'un corpus. Cet outil transforme un jeu de données en site internet, permet de créer des visualisations à partir des données, d'explorer le corpus, et, surtout, d'analyser le corpus grâce à des web services (ou programmes) de fouille de textes [5] (e.g. détection de termes, topic modeling, cf. Cuxac, 2024).

Notre démonstration repose sur un cas d'usage montrant comment répondre à une problématique scientifique en utilisant un corpus Istex. Dans un premier temps, nous expliquons comment interroger l'API pour constituer un corpus en utilisant l'application Istex Search. Dans un second temps, nous présentons de quelle façon explorer et analyser le corpus constitué en utilisant l'outil Lodex. Nous montrons alors comment créer simplement des visualisations à partir du corpus et comment initier une analyse en utilisant les web services de fouille de textes. Cet exposé vise à mettre en lumière le potentiel des outils de l'infrastructure Istex pour rendre l'analyse de corpus scientifiques plus accessible, reproductible et exploitable par la communauté linguistique.

1. L'infrastructure est présentée dans son ensemble sur le site : `https://www.istex.fr/`.
2. Le réservoir Istex est enrichi en continu via les acquisitions pérennes des licences nationales complémentaires aux abonnements courants ou via des projets comme CollEx-Persée (`https://www.collexpersee.eu`).
3. Istex Search est accessible en ligne : `https://search.istex.fr`.
4. Le site du projet (`https://www.lodex.fr/`) présente des exemples de réalisation.
5. Les web services sont tous décrits dans le catalogue Istex TDM : `https://services.istex.fr/`.

Références bibliographiques

Bénard, M., Mestivier, A., Kubler, N., Zhu, L., Bawden, R., De La Clergerie, E., Romary, L., Huguin, M., Nominé, J.-F., Peng, Z., & Yvon, F. (2023). MaTOS : Traduction automatique pour la science ouverte. In F. Boudin, B. Daille, R. Dufour, O. El, M. Houbre, L. Jourdan, & N. Kooli (éd.), *Actes de CORIA-TALN 2023, atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)*, 8-15.

Breucker, P., Cointet, J., Hannud Abdo, A., Orsal, G., de Quatrebarbes, C., Duong, T., Martinez, C., Ospina Delgado, J.P., Medina Zuluaga, L.D., Gómez Peña, D.F., Sánchez Castaño, T.A., Marques da Costa, J., Laglil, H., Villard, L. & Barbier, M. (2016). *CorTexT Manager* (version v2). URL : `https://docs.cortext.net`.

Cuxac, P., Mahut, V., & Niederlender, C. (2017). Istex : Les projets d'enrichissement menés par les équipes de l'Inist. *Workshop Istex*. Nancy, France. `https://doi.org/10.13140/RG.2.2.30707.53281`.

Cuxac, P., & Collignon, A. (2017). Istex, un projet national d'archives documentaires : au-delà de l'accès au texte intégral, l'enrichissement des données par méthodes de fouille de textes. *Analyser la science : les bibliothèques numériques comme objet de recherche in 85e Congrès ACFAS*. Montréal, Canada.

Cuxac, P. (2024). La fouille de textes en IST : Les outils Istex-TDM. *INFORSID ' 24*. Nancy, France.

Delanoé A. & Chavalarias D. (2023). *GarganText, collaborative and decentralized LibreWare*. CNRS, ISC-PIF (UAR3611). GarganText organization on ISCPIFs' GitLab repository : `https://gitlab.iscpif.fr/gargantext/main`.

Gregorio, S., Collignon, A., Parmentier, F., & Thouvenin, N. (2019). Lodex : des données structurées au web sémantique. *Atelier Web des Données de la 19e Conférence sur l'Extraction et la Gestion des Connaissances* (EGC 2019). Metz, France.

Habert, B. (2000). Des corpus représentatifs : De quoi, pour quoi, comment ? In M. Bilger (éd.), *Linguistique sur corpus. Etudes et réflexions—Mireille Bilger*, 11-58. Perpignan : Presses Universitaires de Perpignan.

Hatier, S., & Tutin, A. (2019). Lexique et phraséologie scientifiques transdisciplinaires en sciences humaines : De la modélisation à la création d'une ressource lexicale. *FIU Francophonie et innovation à l'université, Quelle place pour le français scientifique dans un contexte universitaire ?*

Pho, P. D. (2008). Research article abstracts in applied linguistics and educational technology : A study of linguistic realizations of rhetorical structure and authorial stance. *Discourse Studies*, 10(2), 231-250. `https://doi.org/10.1177/1461445607087010`.

Rastier, F. (2004). Enjeux épistémologiques de la linguistique de corpus. In G. Williams (éd.), *La linguistique de corpus*, 31-46. Rennes : Presses Universitaires de Rennes.

Silberztein, M. (2016). *Formalizing Natural Languages : the NooJ approach*. London, Hoboken : Wiley.

SMIQQDA

un logiciel libre pour l'analyse de données (numériques) multimodales

Tiago JOSEPH
Multiples (Language in Society), UGen

Le logiciel **SMIQQDA**, *Social media interface for quantitative and qualitative data analysis,* a été développé dans le cadre d'une thèse de doctorat en analyse du discours sur un corpus texte-image de profils et de publications d'Instagram.

Initialement conçu pour des données de réseaux sociaux, le logiciel SMIQQDA s'adapte également à d'autres corpus multimodaux texte-images (articles de presse, affiches, bande-dessinées, livres illustrés, etc.). Il s'avère particulièrement utile pour l'analyse, l'archivage et le partage de données quantitatives, multimodales et numériques, dans une démarche de méthode mixte.

SMIQQDA répond ainsi au besoin de développer des outils d'analyse mixte et multimodale en sciences sociales, en particulier en linguistique et en sciences de l'information et de la communication. Inscrite dans une perspective de science ouverte, la présentation du logiciel vise à faciliter son utilisation et à le diffuser auprès des communautés de linguistique de corpus et d'humanités numériques, et invite à de futurs développements collaboratifs.

La première version et le code source de SMIQQDA seront prochainement disponibles sur Github et Zenodo[1], gratuitement, sous la licence libre et open source GPL3. C'est le cas également de la documentation technique détaillée, comprenant un manuel (anglais/français). Basé sur les briques logicielles Python (3.11.6) et Flask (2.2.3), le logiciel est exécutable dans n'importe quel environnement (Windows, OS, Linux, etc.) et son interface est en anglais.

Sur base d'un corpus-test composé de trois sous-corpus (profils, publications, OCR des publications), la démonstration présentera en détail les grandes fonctionnalités de SMIQQDA :

- Constituer une base de données multimodales (structurer, synthétiser, articuler différents sous-corpus texte-image) ;
- Éditer les données (vérifier, corriger, mettre à jour, enrichir) ;
- Explorer les données (afficher, chercher, trier, filtrer, revenir au contexte d'origine) ;
- Analyser les données (annoter, catégoriser, réaliser des statistiques textuelles) ;
- Gérer les données (archiver, partager, créer des sous-corpus texte-image) ;
- Articuler plusieurs outils quantitatifs d'analyse de textes (TXM) et d'images (Panoptic), et synthétiser leurs analyses.

SMIQQDA intéressera toute personne travaillant sur des (gros) corpus multimodaux texte-image (voire texte-vidéo), depuis des méthodes mixtes, qualitatives ou quantitatives.

[1] Un *software paper* est en cours d'évaluation.