Dictionnaires et Néologismes

M1 Langue et Informatique

2024-2025

Mini-mémoire

Gaël Lejeune, Sorbonne Université

---

1 Organisation du Travail

Une organisation en binômes vous est proposée mais vous pouvez choisir éventuellement d'y aller seul.

2 Le rendu

Votre rendu sera constitué de 2 parties :

— un **rapport technique** de 5 pages de contenu (minimum) à 10 pages (maximum) présentant la tâche, le jeu de données et les approches utilisées. Du matériel supplémentaire (tableaux de statistiques, figures ...) peut être mis en annexes si nécessaire.

— un **code fonctionnel** permettant de reproduire les résultats de votre chaîne de traitement. Si vous avez besoin d'utiliser un outil externe (par exemple un extracteur de termes), vous devrez préciser dans votre rapport la série exacte d'étapes à suivre pour reproduire votre résultat

Calendrier :

— 27/03/2025 : Présentation des sujets
— 11/04/2025 : Date limite de **Choix du sujet**
— 11/04/2025 : Démarrage du projet
— 11/05/2025 : Dépôt sur Moodle de votre rapport et de votre code
— 13/05/2025 : Dépôt sur Moodle des diapos pour la présentation orale
— 15/05/2025 : **Présentation orale** (10mn de présentation et 5mn de questions)

3 Sujets 2024-2025

3.1 Compléments d'Objet Implicites, L'omission du COD : une explication pragmatique ?

Contexte : En français, certains verbes dits « transitifs », c'est-à-dire sélectionnant un complément d'objet direct (COD), peuvent être employés aussi sans COD. On peut par exemple dire que l'on est en train de manger, de boire, de lire, d'écrire voire de consulter. Alors, pourquoi ne peut-on pas dire de même que l'on est en train de dévorer, d'ingérer, de feuilleter, ou d'inscrire ? Pourquoi cela semble-t-il étrange ou moins naturel ? Plusieurs explications ont été proposées par les linguistes, certaines évoquant l'arbitraire des langues, d'autres s'appuyant sur les caractéristiques lexicales des verbes, d'autres encore sur leur insertion dans des constructions spécifiques. L'hypothèse que nous souhaitons tester ici est d'ordre pragmatique ou communicationnelle : le fait qu'un

verbe transitif décrive une action particulièrement saillante socialement, qu'il évoque une routine bien identifiée par les locuteurs, favoriserait l'omission du COD.

But Peut-on vérifier cette hypothèse en examinant l'omission du COD dans différents corpus plus ou moins spécialisés thématiquement ? En premier lieu, savoir dans quelle mesure on peut repérer automatiquement, et de façon fiable, la présence/absence de COD (avec un étiqueteur syntaxique) et de lister les verbes concernés. Une fois le repérage des omissions effectué, il s'agira de typer (par verbe, par type de construction de phrase ...) ces phénomènes.

Sources et méthodes Un corpus web étiqueté en POS (Web as Corpus) et un corpus non étiqueté (Europresse) pour repérer les verbes concernés et les ordonner.

3.2 Identification automatique de défigements

Contexte : Le défigement est un phénomène linguistique par lequel sont modifiées des séquences figées sur le *continuum* linguistique. Ainsi, une expression telle que « que la force soit avec toi » peut devenir « que la**tri**force soit avec toi » ou encore « que la force **tranquille** soit avec toi ». Afin d'étudier le passage d'une séquence figée à une séquence défigée, nous utilisons la typologie des transformations (ou métaplasmes). Les transformations possibles sont la substitution, l'insertion, la permutation ou la suppression d'un ou de plusieurs tokens et/ou caractères. Cependant, il est parfois difficile de dire si nous sommes en face d'une séquence défigée ou d'une variation acceptable d'une séquence figée (créée par un accord en genre ou en nombre, la conjugaison d'un verbe à un autre temps, ...).

But : À l'aide d'un outil dédié (ASMR), nous cherchons à extraire des défigements candidats pour deux séquences figées : « travailler plus pour gagner plus » et « que la force soit avec toi ». L'outil classe ensuite ces défigements candidats en leur attribuant un score selon plusieurs critères. En analysant les classements produits, pouvons-nous identifier des défigements pour chaque transformation possible ? Les scores produits permettent-ils de faire la différence entre un bon candidat (= un défigement) et un mauvais candidat (= une variation simple, une séquence libre, ...) ?

Sources et méthodes : Le jeu de données est constitué de deux fichiers au format json (un pour chaque expression étudiée). Le but est d'utiliser l'outil ASMR afin de produire des classements de défigements pour ces deux expressions. Ces classements devront ensuite être étudiés afin d'observer si (i) on y observe des défigements, (ii) les scores permettent de différencer les défigements des autres candidats et (iii) les scores s'ordonnent en fonction des transformations observées (substitution, insertion, permutation ou suppression).

3.3 Lexique et débats politiques

Contexte : La communication humaine se base sur la compréhension partagée du sens des mots. C'est en particulier vrai dans le cadre de la vie politique, qui repose en partie sur le fait que le public visé par un orateur comprenne le message qui lui est adressé. Ainsi, de nombreuses études ont cherché à explorer comment comment le sens de certains mots est manipulé, et comment celui-ci peut évoluer au cours du temps. Ces questions ce sont souvent posées sur un corpus anglophone, mais il est également

intéressant de se les poser sur un corpus francophone, et plus précisément concernant la vie politique française.

Buts :

- Le style de discours de certain-e-s hommes ou femmes politiques est-il identifiable ? Peut-on rattacher un style à un parti ?
- Observe-t-on une évolution du sens ou du contexte de certains mots employés dans les débats, et peut-on rattacher certains de ces changements à des évènements de l'actualité de l'époque ?

Sources et Méthodes Le dossier "FREDSum Parliament" [1] accessible sur Ortolang contient des transcriptions de débats ayant pris place à l'Assemblée Nationale et au Sénat entre 2004 et 2024.

3.4 Entités nommées spatiales en contexte bruité

Contexte La transcription est, dans notre cas, le passage d'un document sur un média informatiquement opaque vers un média exploitable numériquement. Ce processus est particulièrement présent en Linguistique, Sciences Humaines et Sociales (LSHS), permettant de rendre exploitable, entre autres, des enregistrements d'entretiens, des documents papiers ou encore des articles issus de pages internet. Le bruit, une résultante de la transcription automatique, qualifie les erreurs générées lors de ce processus ; le bruit peut alors représenter l'ajout, la délétion ou la modification d'un ou plusieurs caractères en regard à la source.

Les outils présentant généralement leurs performances sur des données non bruités, il peut s'avérer complexe d'évaluer et comparer la pertinence des outils pour une tâche donnée sur des corpus bruités. Ainsi, l'objectif de ce projet sera d'évaluer l'impact de différents niveaux de bruits sur les outils de Reconnaissance des Entités Nommées (REN).

But Quel est l'impact du bruit sur la REN ? Quels outils y sont les plus sensibles ? À quels niveaux de bruit ?

Sources et Méthodes Les données exploitées seront celles de WikiNER [2] [**?**], un jeu de données annoté manuellement mais présent en tant que jeu d'entraînement de nombreux outils, veillez à bien en utiliser la partie "test".

Pour les outils : une librarie permet de bruiter artificiellement les textes, `string-noise` [3] ; une autre permet l'évaluation du niveau de bruit, `jiwer` [4] ; et plusieurs permettent l'extraction d'entités nommées (`SpaCy` [5], `flair` [6], `stanza` [7], ...).

1. https://github.com/linagora-labs/FREDSum
2. https://huggingface.co/datasets/mnaguib/WikiNER
3. https://github.com/dleemiller/string-noise
4. https://github.com/jitsi/jiwer
5. https://spacy.io/
6. https://huggingface.co/flair
7. https://stanfordnlp.github.io/stanza/

3.5 Versification et néologismes

Contexte L'analyse des textes versifiés, documents textuels composés uniquement de vers (theâtre, poésie, chanson, ...), passe souvent par une analyse métrique. Cependant, peu d'outils existent pour le français.

De plus, parmi les outils existants, nombre d'entre eux reposent sur des lexiques, faisant correspondre les formes connues à une division syllabique. D'autres, quant à eux, reposent sur des systèmes à base de règles ou des réseaux neuronaux. On s'attend alors à ce que un système basé sur des lexiques patisse de la présence de termes hors-lexique, tels que les néologismes, mais qu'en est-il vraiment ? Quel est l'impact de la présence de néologismes sur les autres méthodes ?

Le projet consistera à évaluer la présence de néologismes au sein de vers, puis évaluer l'impact de leur présence sur différents outils de métrique automatique.

But Quelle est l'influence des néologismes sur la mesure des vers ? Quels sont les outils/types d'outils les plus impactés ? Comment pallier la présence de néologismes sur la mesure de vers ?

Sources et méthodes Le corpus fourni sera issu de la base Métrique en ligne [**?**] et sera composé d'une suite de couples vers / métriques.

3.6 Le lexique autour des lieux de Paris

Contexte : La chanson représente une partie importante de la culture populaire. L'évolution des mentalités au sein d'une société peut ainsi être mesurée à l'aune des thèmes qui sont abordés selon les périodes temporelles. L'objectif de ce projet, orienté linguistique de corpus, est d'analyser dans quelle mesure les noms de lieux évoqués dans la chanson française ont évolué du début du XXème siècle jusqu'à aujourd'hui. Pour des raisons pratiques, le sujets est restreint aux lieux parisiens (monuments, bâtiments, rues ...) car Paris est le lieu le plus mentionné dans la chanson française.

But : Identifier si la gentrification de Paris (= éloignement des classes populaires) se traduit dans la culture populaire à travers les noms de lieux de Paris et leur contexte d'apparition.

Sources et Méthodes Un corpus généraliste de 30.000 chansons en français ainsi qu'un corpus spécifique sur le rap, un ressource de noms de Paris pour filtrer les entités nommées correspondant à des lieux de Paris.

3.7 La liberté d'expression dans la presse

Contexte : Le projet s'inscrit dans le cadre d'une recherche doctorale visant à comprendre comment la liberté d'expression, émerge et se structure dans le débat public. L'étude s'appuie sur un corpus d'environ 18 000 articles de presse issus d'Europresse, publiés entre 2000 et 2024, contenant le mot-clé "liberté d'expression". L'objectif est de distinguer les articles où la liberté d'expression est le sujet principal, structurant le contenu, de ceux où elle n'est qu'un élément secondaire, mentionné sans être central. L'objectif est de distinguer ces deux usages pour ensuite pouvoir étudier la centralité de la notion en fonction des thèmes/sujets traités.

But :

Identifier les articles où la liberté d'expression est le sujet principal et ceux où elle n'est qu'une mention secondaire en attribuant à chaque article un "statut" ("principal" ou "secondaire") ou un score de centralité, en fonction des possibilités. Intégrer ces résultats sous forme de métadonnées exploitables dans le corpus.

Sources et méthodes : Le corpus est composé d'articles de presse au format .txt, .csv, ou .xml, enrichis de métadonnées. Un sous-corpus pourra être fourni pour les analyses initiales. Il s'agira d'exploiter les méthodes d'aznalyse diachronique vues en cours

3.8 Comment caractériser la langue d'Annie Ernaux ?

Contexte Contexte : Prix Nobel de littérature en 2022, Annie Ernaux est l'autrice d'une vingtaine de récits à caractère autobiographique, ce qui confère à son œuvre une certaine homogénéité thématique. Toutefois, du côté de son style, l'autrice revendique un tournant esthétique à partir de sa quatrième œuvre – ce à quoi la critique journalistique comme universitaire acquiesce le plus souvent. On distinguerait donc, d'une part, les trois premiers ouvrages d'Annie Ernaux [8] écrits selon un principe de « dérision » [9], riches syntaxiquement [10] figuralement [11], et lexicalement [12]. D'autre part, on trouverait le reste de l'œuvre de l'autrice à partir de La Place, rédigé au moyen de sa désormais célèbre « écriture plate ». À rebours de la langue révoltée de ses premiers écrits, « l'écriture plate » constituerait un refus d'esthétisation ; il s'agirait d'écrire au ras du réel, sans verser du côté du misérabilisme ou du côté du pittoresque.

But : Comparaison interne, est-il si certain que l'œuvre de l'autrice soit absolument duale ? Peut-on repérer des éléments (syntaxiques, sémantiques, figuraux) qui se retrouveraient au fur et à mesure des ouvrages, et qui résisteraient à la césure instaurée par La Place ? La différence est-elle nette entre les publications du début et les plus récentes d'Annie Ernaux ? Doit-on minorer, nuancer, ou au contraire valider la revendication, et l'analyse que fait l'autrice de son propre style ?

Comparaison externe, est-il possible d'identifier les attributs spécifiques de l' « écriture plate » (propositions juxtaposées, vocabulaire restreint par exemple) d'Annie Ernaux en regard d'autres esthétiques contemporaines ? Est-ce que le (non) style d'Annie Ernaux peut être mis en relation avec celui d'autres auteurs et autrices écrivant à la même époque ? En d'autres termes, est-ce une écriture propre à Annie Ernaux, ou peut-on dégager ici un trait générationnel, voire un fait d'époque (une mode) ?

Sources et méthodes Un corpus électronique des œuvres d'Annie Ernaux, pour la comparaison externe il faudra constituer un corpus de contraste (Marguerite Duras, Michel Houellebecq, Virginie Despentes, Maelys de Kerangal par exemple).

8. Les Armoires vides, Ce qu'ils disent ou rien et La Femme gelée
9. Selon le mot de l'autrice
10. Propositions subordonnées, clivées(c'est une pomme que je mange), pseudo-clivées(ce que je mange, c'est une pomme)
11. Anacoluthes, aposiopèses, répétitions, accumulations, énumérations etc.
12. Mots aux connotations populaires, régionalismes (patois normand)

3.9 Terminologie dans le domaine de l'écologie : l'exemple de la dégradation des terres

Contexte : Il existe plusieurs définitions de la dégradation des terres, notamment dans les rapports des grandes institutions internationales liées au climat et à l'environnement. Par exemple, cela s'exprime par « [le déclin] de l'utilité des terres, de la biodiversité, de la fertilité des sols et de la santé globale » (UNCCD) ou par « le déclin ou la perte de la biodiversité, des fonctions des écosystèmes ou de leurs avantages pour les populations » (IPBES). Ces définitions institutionnelles sont parfois utilisées dans des publications scientifiques cherchant à caractériser la dégradation des terres. Cependant, d'autres études utilisent leurs propres critères pour déterminer s'il y a ou non dégradation des terres, par exemple en se focalisant sur l'évolution du couvert végétal suivi par des méthodes de télédétection. Bien que les principaux discours sur la dégradation des terres déclenchent d'importants financements pour des mégaprojets comme la Grande Muraille Verte au Sahel, ce flou dans sa définition interroge sur la cohérence entre l'étendue et la gravité de la dégradation des terres selon ces grandes institutions et selon la littérature académique.

But : Recenser les définitions existantes pour la dégradation des terres, dans 2 corpus (institutionnel et académique), et à retracer l'évolution de leur usage (notamment, la perméabilité entre ces 2 corpus, la prédominance de certaines définitions et leur circulation entre les documents et entre les auteurs).

Sources et méthodes : Deux corpus de documents, l'un institutionnel et l'autre, scientifique, qui permettront de caractériser des modifications dans les termes des définitions.

3.10 Statut des « emprunts » en français

Contexte : La globalisation et l'évolution des technologies de la communication favorisent les effets du contact des langues dans les pratiques langagières actuelles. Cela se manifeste par la présence d'emprunts à différentes langues selon des modalités de contact différentes : contact avec des langues de l'immigration ou par des interactions entre locuteurs. Il peut s'agir aussi du contact avec des langues par la familiarité avec des référents culturels. On pense par exemple aux emprunts à l'anglais qui, pour de nombreuses langues, sont le résultat de la familiarité des locuteurs avec la musique ou le cinéma anglosaxon. Les mouvements migratoires, le brassage des langues et des cultures, en particulier dans les grandes métropoles, et la facilité des communications internationales conduisent à porter un regard renouvelé sur les emprunts : certains relèvent de l'alternance codique dans des échanges entre locuteurs plurilingues et d'autres relèvent de l'intégration d'éléments d'une langue étrangère pour enrichir le vocabulaire de la langue dans une dynamique néologique. Certains mots peuvent avoir ces deux statuts au sein d'une communauté de locuteurs, par exemple, « wesh » dans la pratique langagière en français de locuteurs maitrisant l'arabe peut être utilisé comme une particule interrogative comme en arabe (alternance codique) et comme marqueur discursif, fonction pragmatique que le mot a en français.

But : L'objectif est de repérer quand l'emprunt à des langues étrangères relève de pratiques langagières plurilingues dans lesquelles on observe de l'alternance codique et quand la langue étrangère constitue une ressource (au même titre que d'autres

procédés) pour enrichir le système linguistique d'une nouvelle unité. Cela devrait conduire à tenir compte du cotexte (environnement linguistique dans lequel apparait l'emprunt) et du contexte (recours aux métadonnées).

Sources et méthodes : Le travail se fera à partir du corpus Multicultural Paris French, disponible sur Ortolang. La plateforme donne accès aux enregistrements audios, aux transcriptions faites sous PRAAT et converties en .txt et à des fiches métadonnées qualitatives.