EXEMPLIFICATIONS AVEC ET SANS MARQUEURS : ÉTUDE DE L'ORAL ET DE L'ÉCRIT

par Élodie Esteves, Yidi Huang, Juliette Massy & Aliénor Sauty de Chalon

Dans cette étude, nous nous sommes intéressés au procédé de reformulation dans sa dimension orale et écrite et l'avons étudié grâce à des méthodes du TAL afin de générer des statistiques sur le phénomène et d'en tirer des conclusions sur les différences existantes selon le contexte d'émission. Dans une première partie, nous nous attacherons à définir notre sujet, l'analyse que nous souhaitons en faire et les notions abordées puis nous présenterons le corpus utilisé et la méthode choisie pour l'exploiter. Enfin, nous présenterons nos études statistiques, complétées d'une analyse approfondie de celle-ci.

I. Reformulations et exemplifications, définition et mise en perspective :

Caractéristique naturelle du langage humain, la reformulation est ici considérée dans son sens large comme "un procédé de modification dans le discours d'un segment source par un segment reformulé"[1]. Ce procédé peut être modélisé comme suit et possède plusieurs caractéristiques :

S1 (MARQUEUR | PHASE D'ÉDITION) S2

- un segment source noté S1 qui est le segment que le locuteur souhaite modifier.
- une « phase d'édition » ou un marqueur de reformulation qui peuvent être absents, directs avec un marqueur de reformulation (MR) ou indirects à l'oral (disfluences ou marqueurs discursifs).
- un segment reformulé S2,
- un lien sous-jacent existant entre les deux segments reformulés.

Notons que les reformulations diffèrent selon les modalités : écrite ou orale. Ainsi pour la phase d'édition, l'utilisation de marqueur indirect est dépendante de la

[1] Eshkol-Taravella, I., & Grabar, N. (2018). Reformulations avec et sans marqueurs : étude de trois entretiens de l'oral. *SHS Web of Conferences*, 46, 11003. https://doi.org/10.1051/shsconf/20184611003

modalité orale. L'usage du discours spontané conditionne ainsi la présence des disfluences ou de marqueurs discursifs.

En ce sens, il faut également souligner que le marqueur n'est pas le critère déterminant pour la présence d'une reformulation, celui-ci pouvant être absent. C'est le lien sous-jacent existant entre les segments qui permet de reconnaître le procédé. Le lien sous-jacent étant la partie invariante qui permet d'établir une relation entre les deux segments.

Soulignons ainsi que la reformulation recouvre non seulement la paraphrase mais aussi l'exemplification, la conclusion, la précision, etc…

Dans ce projet, notre attention s'est portée sur un type précis de reformulation, les exemplifications. Comme son nom l'indique, une exemplification correspond à la reprise d'un énoncé antérieur par un second segment contenant des exemples.

1) tu peu rien faire même pas ramasser un simple stylo tomber par terre
2) avec euh ce qu'il faut euh une douche euh sur place

Voici deux exemples d'exemplifications tirés de nos corpus. Le premier est tiré des documents écrits et le deuxième de ceux de l'oral.

Dans les deux exemples, une affirmation correspondant au premier segment est illustrée par un exemple plus spécifique.

Les deux énoncés présentent une syntaxe différente : l'énoncé écrit (1) ne comporte pas de marqueur de reformulation et l'énoncé oral (2) contient une disfluence, la marque d'hésitation *euh*.

Ce sont à ces types de propriétés et de différences que nous nous sommes intéressés dans cette étude. Nous allons ainsi chercher à savoir en quoi l'étude statistique comparative des exemplifications révèle des tendances selon les contextes d'émission. Le contexte d'émission renvoyant à oral ou écrit, discours professionnel, spontané…

II. Données et corpus : annotation XML et prétraitement

1) Nature des données :

Le corpus d'étude est composé de deux sous corpus :

- un corpus oral composé de trois entretiens entre locuteurs témoins et chercheurs, basé sur des questions sur la vie des locuteur à Orléans,
- et un corpus écrit composé de **deux fils de messages** postés sur des forums de santé : l'un étant composé de messages postés par des patients sur un forum consacré aux douleurs de dos et l'autre étant les réponses de médecin sur le forum MaSante.net.

Le corpus était déjà annoté au format XML.

2) Correction de l'annotation :

Ces corpus étaient initialement annotés avec des balises XML. L'annotation portrait sur les reformulations, quelque soit leur caractéristiques.

Seulement, l'annotation contenait de nombreuses erreurs et nous l'avons donc reprise et corrigée. Cette correction concerne les 5 documents du corpus d'étude.

Les corrections apportées sont les suivantes :

- ajout de balises manquantes (fermantes ou ouvrantes) et création d'éléments <reformulation> pour les corpus de l'oraux,
- ajout de guillemets manquants pour les attributs,
- modification des noms de balises quand un couple de balises ne correspondait pas (les erreurs provenant sans doute de fautes de frappes)...

Les corrections ont été rapidement effectuées grâce à des regex et un logiciel de traitement textuel. De plus, nous avons décidé de créer un corpus XML qui serait exploitable automatiquement avec des modules python dédiés et nous avons donc modifié le corpus pour qu'il soit valide dans l'éditeur Oxygen. Les modification consistent en :

- l'ajout d'un élément racine </body>,
- la modification du nom de l'attribut quand une balise possédait deux attributs identiques (par exemple quand deux attributs *rel_pragm* étaient spécifiés le second a été nommé *rel_pragm2*)
- l'ajout d'attributs *mod* et *corpus* aux balises <reformulation> qui permettent de distinguer la modalité (*mod)* orale/écrite des énoncés et leur corpus d'origine (eslo55, eslo33, eslo5, masante, patient). Ceci a été utile lors de la création d'un fichier corpus unique contenant toutes les reformulations extraites automatiquement.

3) Modèle d'annotation XML des données :

L'annotation et les conventions ont été transmises par Madame Taravella. Nous avions à notre disposition deux documents : la typologie des marqueurs et la convention de balisage.

Les segments ont deux catégories syntaxiques. Il y a les catégories syntaxiques avec modifieurs et les catégories syntaxiques seules.

Les catégories syntaxiques avec modifieurs sont des groupes prépositionnels PP, des groupes nominaux PN, des groupes adjectivaux AP et des groupes verbaux VP. Les catégories syntaxiques seules sont les suivantes : les adjectifs (A) , les adverbes (ADV), les noms (N), les présentateurs (PRES), les pronoms (pr), les gérondifs (Ger), les pronoms (P). Les segments sont par ailleurs numérotés au niveau de la balise indiquant leur catégorie syntaxique.

En ce qui concerne les marqueurs, ils sont notés<MR> et quand il n'y en a pas, ils ont une balise auto-fermante <MR/>.

Les marqueurs de reformulation sont notés <MRE>. Ils s'inscrivent dans une typologie. Il y a les marqueurs de dénomination qui sont notés <MDENOM>. Ils sont employés pour les verbes comme "appeler", "nommer". Les autres marqueurs de dénomination sont des marqueurs de désignation notés <MRDESIGN> employés pour des expressions avec "désigner", "signifier", "vouloir dire". Les <MRP> sont les marqueurs de paraphrase, les <MRCOONC> les marqueurs concluants, les <MRCOO>, les marqueurs de coordination, les <MRCOR> les marqueurs de correction.

Dans le corpus oral, il arrive souvent de trouver des marqueurs de disfluence. Les <DA> correspondent aux amorces, les <DH> aux hésitations, les <DI> aux interjections, les <DMD> aux marqueurs discursifs.

Les relations entre le premier segment S1 et le deuxième segment S2 sont notées sous forme d'**attributs** dans le deuxième segment. Il peut y avoir un troisième segment, voire un quatrième, mais cela demeure assez rare et concerne essentiellement le corpus oral dans lequel les énumérations sont fréquentes; le locuteur produit plusieurs reformulations car il cherche ses mots.

Les attributs du deuxième segment sont de deux types: les **relations** et les **modifications**.

- Les relations sont **lexicales**: antonymie/ synonymie; méronymie (relation partie/tout); instance (noms propres, entités nommées); hyperonymie ("un mot dont le sens inclut celui d'autres mots plus spécifiques"; par exemple animal est l'hyperonyme de chien) / hyponymie ("un mot dont le sens est inclu dans celui d'un autre générique"; par exemple mouche est l'hyponyme d'insecte).
- Les relations sont aussi **pragmatiques** : définition/ dénomination/ explicitation/ justification/ précision/ résultat/ paraphrase/ opposition/ **exemplification**.

- Les modifications sont **morphologiques** : la flexion (pour les verbes) et la dérivation (pour les noms).
- Les modifications sont aussi **syntaxiques**. Dans ce cas, elles ne portent que la valeur : passif/actif.

III. Extraction automatique des reformulations avec python, regex et etree :

Dans le cadre de notre étude de corpus, nous avons entrepris différentes analyses statistiques sur les marqueurs utilisés, les modifications morphologiques, ainsi que sur les relations lexicales présentes. Pour ce faire, nous avons développé des scripts Python afin d'analyser automatiquement le corpus et de générer des diagrammes illustratifs.

La première étape de notre démarche consistait à créer un script permettant d'extraire toutes les reformulations, afin d'obtenir un fichier ne contenant que les données pertinentes. Pour ce faire, nous avons utilisé le module regex de Python afin d'extraire

le contenu de chaque élément <reformulation> du corpus. Chaque élément était ensuite enregistré dans un fichier XML respectant une structure préétablie.

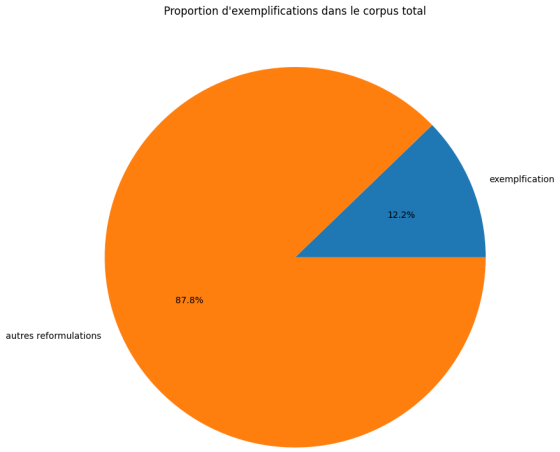
Une fois les reformulations extraites, nous avons développé deux scripts, chacun étant dédié à l'un des angles d'analyse que nous avions choisis : écrit vs oral, et écrit patient vs écrit médecin vs oral. Ces scripts nous ont permis d'effectuer des analyses statistiques sur les éléments pertinents du corpus. Ils

Nous avons réalisé des calculs statistiques et des analyses approfondies en utilisant des scripts Python pour étudier le corpus, extraire les reformulations et mener nos investigations selon les angles d'analyse sélectionnés.

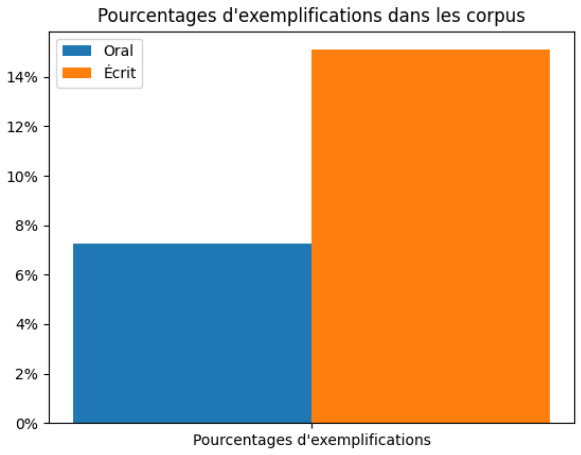
IV. Analyse :

1) Première approche : écrit vs oral :

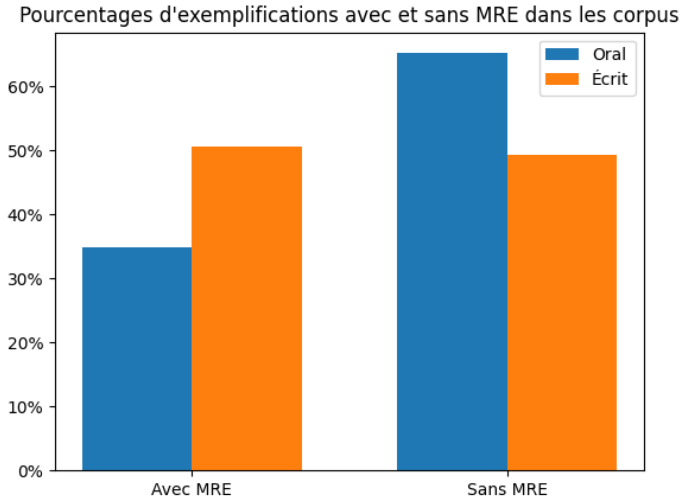
Dans notre corpus, les exemplifications sont au nombre de 108. Elles représentent 12,2% des reformulations du corpus.



En abordant notre corpus sous notre premier angle, on peut voir que 15% des reformulations de l'écrit sont des exemplifications contre 7% à l'oral.



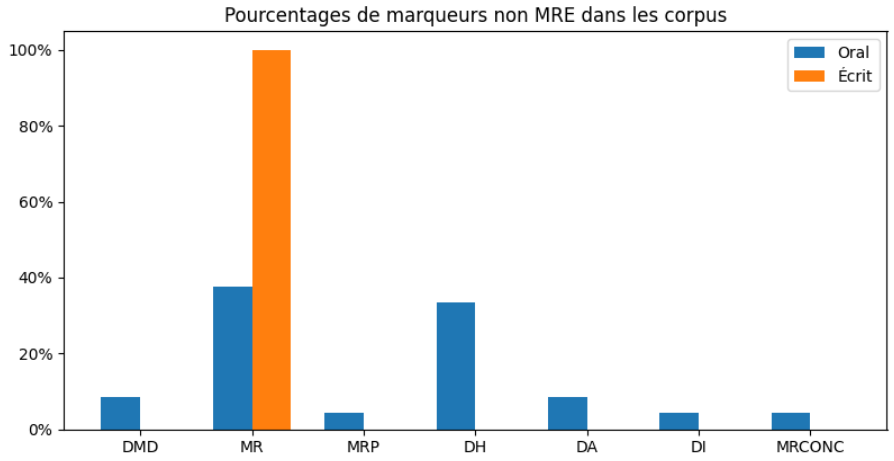
Parmi ces exemplifications, plusieurs types de constructions existent et ce notamment à cause de la présence ou l'absence de MRE.



À l'écrit, nous pouvons voir qu'il y a à peu près autant d'exemplifications avec et sans MRE, un peu plus avec, contrairement à l'oral où seules 34% des exemplifications possèdent des marqueurs de type MRE.

À l'oral, nous retrouvons notamment MRE composé d'expressions courtes, un seul mot, et le marqueur d'exemplification par excellence *par exemple*. À l'écrit, nous pouvons observer une plus grande diversité des marqueurs et des marqueurs plus longs même si certains se retrouvent dans les deux modalités. **La question qui se pose alors est de savoir la nature des marqueurs qui sont utilisés à la place de <MRE>** et de savoir si celle-ci dépend de la nature du corpus. L'hypothèse que l'on

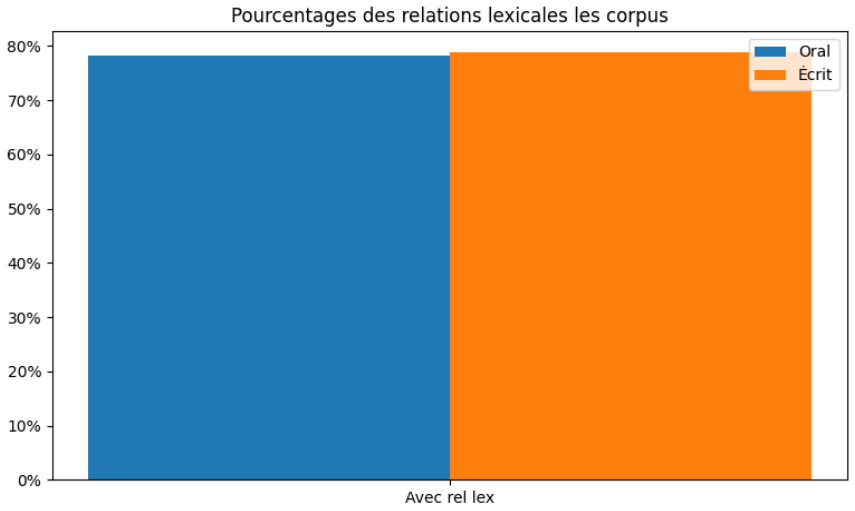
peut formuler sur cette répartition est qu'il y a dans les corpus oraux des éléments introducteurs qui ne sont pas des marqueurs comme les disfluences



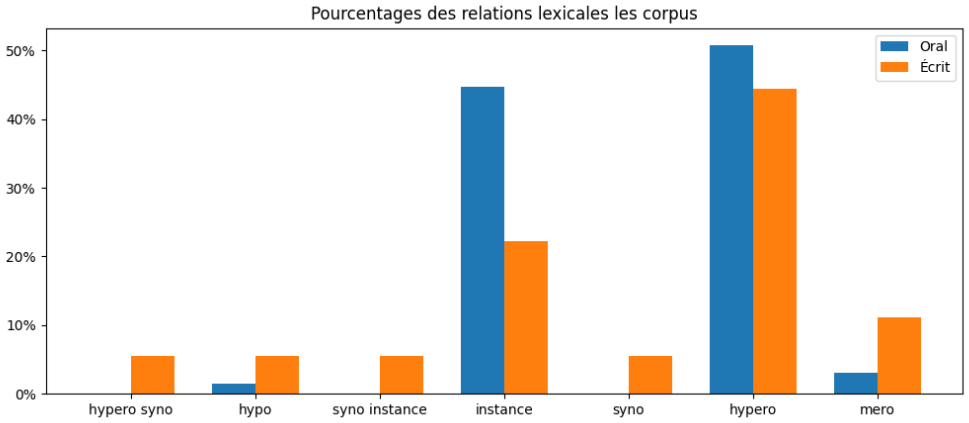
À l'écrit, les <MR> simples sont utilisés exclusivement comme autre marqueurs dans les exemplifications et correspondent surtout à de la ponctuation comme les deux points, les parenthèses, les virgules, etc... À l'oral, l'hypothèse est validée : les marqueurs sont répartis dans plusieurs classes de disfluences et notamment les disfluences d'hésitations. Notons ici aussi, des marqueurs classiques de type présentatif comme *c'est*.

Maintenant que nous avons étudié les marqueurs explicites de reformulations, intéressons-nous aux relations qu'entretiennent les segments entre eux.

Les liens lexicaux sont présents dans 79 % des exemplifications et dans un pourcentage équivalent dans chaque corpus, comme nous pouvons le voir dans le diagramme ci-dessous.



Concernant la nature de ces liens, nous pouvons voir ci dessous, qu'une exemplification peut comporter plusieurs liens lexicaux, c'est-à-dire que plusieurs unités peuvent être liées lexicalement. C'est le cas pour 10% des exemplifications de l'écrit avec les relations d'hyperonymie/synonymie et celles de synonymie/instance. Une tendance globale se détache de l'analyse statistique, celle de la forte présence des relations d'hyperonymie, la plus présente dans les deux corpus, et d'instance. On verra plus tard que ces tendances peuvent s'expliquer, notamment à l'écrit, à cause de la nature des forums étudiés.



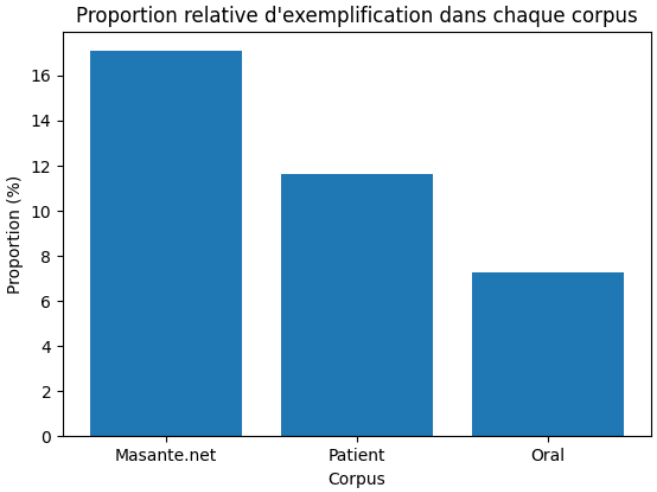
Concernant le deuxième type d'attribut que l'on peut trouver sur le second segment, celui des modifications morphologiques, celui ci n'apparait que dans le corpus oral. Il est ainsi présent dans 17% des exemplifications. Nous pouvons souligner, qu'il est

récurrent à l'oral d'avoir à corriger sa formulation ou de reprendre ce que l'on a dit pour en changer sa forme alors qu'à l'écrit, les énoncés ne sont pas écrits avec de tels reprises (si ce n'est peut être pour un aspect stylistique ou dans un langage très proche de l'oral). De plus, les modifications de ce type sont réparties de manière égale dans notre corpus entre flexion et dérivation (50/50).

Nous avons donc constaté des variations entre écrit et oral, notamment au niveau des marqueurs. Dans l'objectif d'approfondir notre étude, nous avons ensuite eu recours à une seconde approche en séparant écrit des médecin et ceux des patients. La question est de savoir si une différence se dégage à ce niveau là et si l'un se rapproche ou non de l'oral.

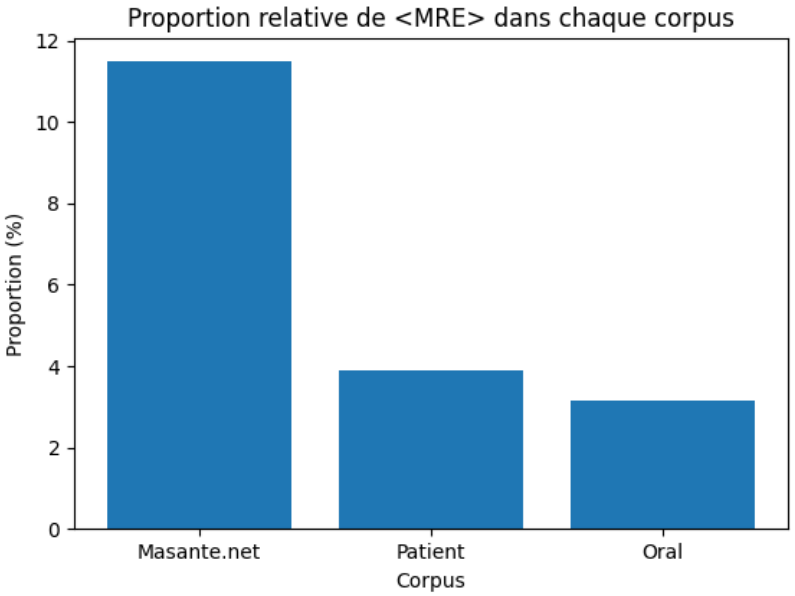
2) Seconde approche : écrit patient vs écrit médecin vs oral :

Au niveau de la proportion d'exemplification, une différence est déjà visible. Le corpus médecin *masanté* possède ainsi une proportion de 17% d'exemplifications pour toutes ces reformulations. On peut penser que cela est fortement lié à la nature, au contexte du discours. Pour expliquer à des patients qui n'ont pas forcément des connaissances scientifiques spécifiques ou pour préciser leur propos, les médecins ont souvent recours à des exemples et cela constitue par ailleurs une particularité de leur discours avec des non-professionnels.



Concernant l'utilisation des marqueurs, comme on peut le voir ci-dessous, les résultats pondérés révèlent que le corpus écrit par les médecins présente une proportion beaucoup plus importante de MRE (11%) que les deux autres corpus qui présentent par ailleurs des proportions similaires (3%). Cette observation nous a conduit à

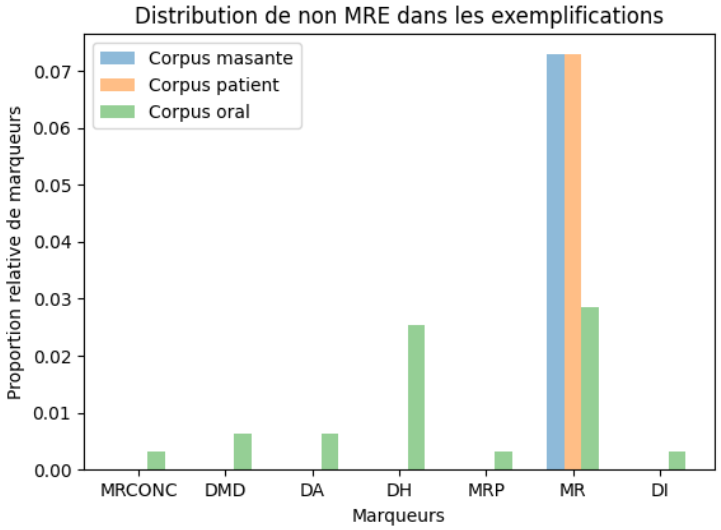
formuler l'hypothèse suivante : le corpus spécifique aux douleurs de dos est plus proche de l'oral car les patients utilisent un langage plus familier lorsqu'ils communiquent sur Internet avec d'autres patients.



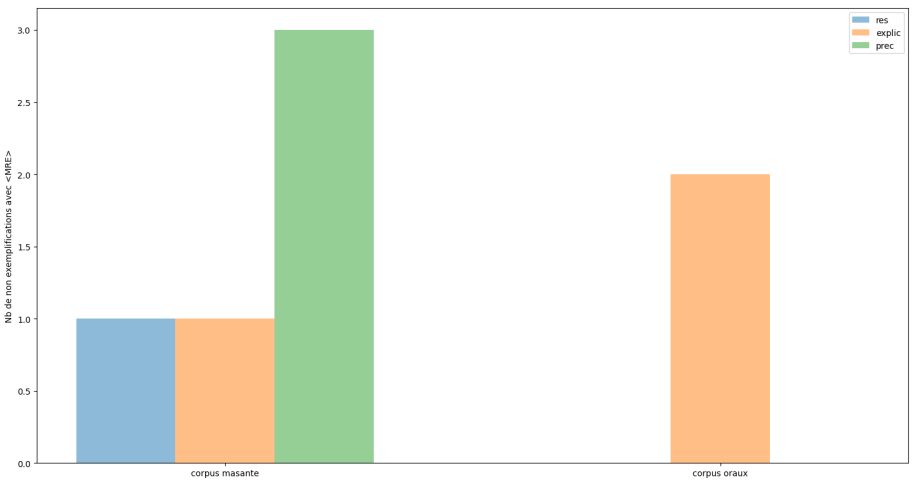
À partir de l'analyse des dictionnaires sur les contenus des exemplifications, il est observé que les marqueurs d'exemplification sont plus variés dans le corpus du forum *masante*. Ce corpus présente une utilisation plus fréquente de marqueurs spécifiques ». En revanche, certains marqueurs, tels que « par exemple » sont plus fréquents dans le corpus oral. Ceci s'explique logiquement par le fait que ces marqueurs sont plus couramment utilisés dans le discours quotidien pour introduire des exemples. De plus, il est clair que le forum dédié aux patients souffrant de douleurs de dos présente des similarités avec le corpus oral, tant en termes de nombre que de diversité des marqueurs d'exemplification. Ces observations confirment ainsi notre hypothèse précédente.

Ici aussi, l'étape logique suivante a été de regarder les marqueurs non MRE dans les corpus.

Nous avons constaté, conformément à nos analyses précédentes, que les marqueurs de disfluences sont fréquemment utilisés pour l'exemplification dans le corpus oral, car c'est plus naturel dans le discours verbal. De plus, il existe d'autres types de marqueurs que nous allons examiner plus en détail pour mieux les comprendre : dans le corpus patient, ce sont plutôt des ponctuations, notamment les deux points, qui jouent le rôle d'introducteur des exemples.

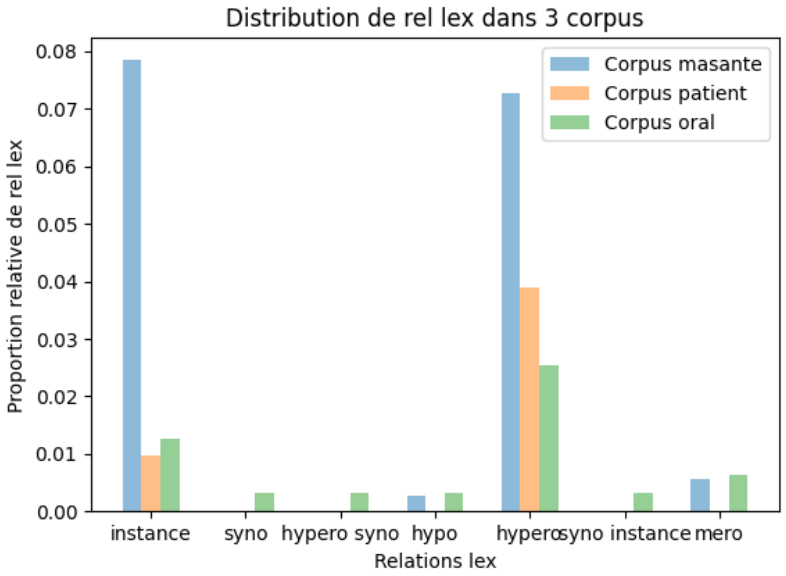


En plus des exemplifications sans MRE, une autre étude est réalisée sur l'utilisation des MRE dans d'autres types de relations pragmatiques. Les MRE peuvent également être utilisés hors exemplification pour introduire un résultat, une précision ou une explicitation. Notons que pour ces types de reformulations, qui sont minoritaires, les MRE sont toujours employés en combinaison avec d'autres types de marqueurs.



Finalement, l'étude sur les relations lexicales pour l'exemplification dans 3 corpus démontre que les médecins adoptent une approche où ils utilisent une majorité

d’hyperonymes pour expliquer des concepts spécialisés, tout en fournissant des instances pour illustrer et clarifier les informations médicales. Cela reflète leur volonté de rendre les explications plus accessibles et compréhensibles pour les patients, en utilisant des termes et des exemples adaptés à leur niveau de connaissance et à leur compréhension.



V. Conclusion :

Notre étude vise à analyser les ewxemplifications à l’oral et à l’écrit en examinant les corpus du forum “MaSanté”, du forum “Douleur de Dos” ainsi que des corpus oraux. Une analyse approfondie de ces corpus met en évidence une diversité de marqueurs et des relations multidimensionnelles entre les segments, notamment sur les plans sémantico-pragmatiques, lexicaux et morphologiques.

Tout d’abord, nous avons constaté que les exemplifications constituent une sous-catégorie des reformulations, souvent marquées par un marqueur spécifique <MRE>, mais également introduites par d’autres types de marqueurs, voire sans marqueur du tout. De plus, les modalités de communication semblent influencer la réalisation des exemplifications, avec la présence de disfluences à l’oral ou l’utilisation de signes typographiques à l’écrit.

En approfondissant l'analyse des corpus et en comparant la variété et le nombre de marqueurs ainsi que les relations existantes, il est possible de constater une similarité entre le corpus "Douleur de Dos", composé de messages rédigés par des patients pour partager leurs expériences et chercher des solutions, et le registre oral en ce qui concerne la formulation et le vocabulaire utilisés.

En plus, les corpus oraux et non professionnels présentent une diminution du nombre et de la diversité des marqueurs d'exemplification, car d'autres marqueurs prennent en charge leur rôle. Par exemple, les forums non professionnels utilisent des ponctuations pour introduire des exemples dans un langage familier, tandis que les corpus oraux utilisent des marqueurs de disfluences tels que des amorces, des interjections et des hésitations, ce qui est naturel, car la communication verbale est plus libre et spontanée.

En revanche, le forum dédié aux médecins présente une utilisation importante des marqueurs de reformulation. Cela s'explique par le fait que les médecins, en tant que professionnels, utilisent fréquemment des terminologies médicales qu'ils doivent rendre compréhensibles pour les patients. Ainsi, ils ont recours aux exemples pour expliquer ces termes spécialisés et clarifier les traitements professionnels aux non-experts.

En résumé, à travers les analyses réalisées, l'analyse des fréquences des marqueurs d'exemplification pourrait fournir des indications sur le profil des locuteurs, allant des professionnels aux non-professionnels, en comparant l'utilisation des marqueurs d'exemplification dans les contextes oral et écrit, ainsi que dans les échanges entre patients et médecins.