Université Toulouse Jean Jaurès II

Département de Sciences du Langage

Master Linguistique, Informatique et Technologies du Langage (LITL)

Mémoire de Master 1

Étude de l’emploi des collocations dans des phrases produites par l’humain et dans des phrases générées par système d’IA

Ethan TOURON

Sous la direction de Madame Cécile FABRE et Monsieur Ludovic TANGUY

Juin 2026

Remerciements

Je souhaite avant tout remercier sincèrement mes encadrants, Madame Cécile Fabre et Monsieur Ludovic Tanguy, pour leur accompagnement tout au long de ce projet passionnant. Je suis reconnaissant du temps qu'ils m'ont accordé, de leurs commentaires et de leurs retours qui m'ont permis d'avancer et de progresser, et qui m'ont encouragé à toujours pousser ma réflexion plus loin.

Merci à l'ensemble des enseignants du Master LITL et du Master LiCo dont les enseignements ont contribué à alimenter mon intérêt pour la linguistique une année supplémentaire en me permettant d'enrichir mes connaissances sur ce vaste domaine.

Je souhaite remercier mes camarades de M1 pour leur bienveillance et l'agréable climat de travail qui a été construit au sein de notre promotion.

Je remercie mes parents pour leur soutien. Merci à ma mère pour son écoute, ses conseils, son partage de connaissances et d'anecdotes sur le monde de la recherche et pour son précieux travail de déchiffrage.

Je tiens à remercier Sofia et Bastien pour leur présence, leur soutien constant et leur compréhension tout au long de la rédaction de ce mémoire. Je leur demande pardon pour l'ensemble des soirées-voc que j'ai déclinées au profit de ce travail.

Je souhaite également remercier Maxime pour nos retrouvailles occasionnelles et pour sa positivité.

Merci à Lionel pour ses encouragements qui m'ont permis d'aborder la dernière phase de ce travail avec confiance.

Je salue Sangho et le remercie pour son enseignement du coréen qui m'a apporté du réconfort tout au long de cette année. Je voudrais également le remercier pour nos échanges captivants autour de la linguistique et son intérêt pour mon sujet de mémoire. Je lui souhaite bon courage pour la finalisation de sa thèse.

Enfin, merci à Monsieur Constantinos Raïos qui m'a permis de garder un pied à Thessalonique tout en étudiant à Toulouse et pour son aide étymologique.

Table des matières

Index des figures

Index des tableaux

Introduction

Aujourd'hui, les systèmes d'Intelligence Artificielle (IA) – et plus précisément les *Large Language Models* (LLMs) – sont largement employés par le grand public pour leur faculté à produire des textes linguistiquement et conceptuellement cohérents. Malgré la nature purement probabiliste de ces systèmes, la qualité des productions générées en réponse aux prompts explique leur popularité auprès des utilisateurs. Ce sujet d'actualité suscite alors de nombreux questionnements quant aux capacités linguistiques démontrées par les modèles. Plusieurs études récentes (dont celles d'Alavoine et al., 2024 et Muñoz-Ortiz et al., 2024) ont comparé des textes produits par ces systèmes avec des textes rédigés par des humains afin d'en déceler les différences linguistiques globales. Aussi, des études ont été menées sur l'aptitude des LLMs à identifier certains objets complexes du langage naturel. Les auteurs Hashiloni et al. (2025) se sont précisément intéressés aux expressions idiomatiques et aux *multi-word expressions* (MWEs), démontrant que ce type de séquences peut présenter un challenge pour les LLMs.

S'inscrivant dans la continuité de ces travaux, ce mémoire a pour objectif de comparer des phrases produites par l'humain (que l'on nommera *phrases sollicitées*) avec des phrases produites par système d'IA (*phrases artificielles*) pour une même tâche en français. Ces phrases seront comparées selon des perspectives structurales, lexicales et morpho-syntaxiques, puis de manière ciblée sur la réalisation de certaines séquences du langage naturel.

Notre matériel d'étude a été constitué à partir de données issues du projet *Evolex*, dont l'objectif premier est d'étudier la fluence et l'accès lexical. Les données *Evolex* (Gaume et al., 2018) consistent d'abord en un jeu de paires de mots issues d'une tâche d'association lexicale avec amorçage. Ces paires rassemblent deux noms susceptibles d'entretenir différents types de relations lexico-sémantiques (ex : *Maladie – Guérison* : relation d'antonymie). Une seconde partie des données issues du même projet regroupe des phrases rédigées par des sujets humains à partir de paires de mots *Evolex*. Ces phrases résultent ainsi d'une tâche de production de phrases avec mots imposés (ex : paire *Sac – Dos* : « Je porte un sac à dos. »). L'objectif du présent mémoire résidant dans la comparaison de phrases sollicitées et artificielles, nous avons récolté les phrases produites dans le cadre du projet *Evolex* afin d'obtenir un premier jeu de données sollicitées. Il convenait ensuite de construire un second jeu de phrases artificielles équivalent à partir des mêmes paires de mots en utilisant un système d'IA.

Plusieurs questions ont émergé de la mise en perspective de la littérature avec des observations préliminaires conduites sur les deux jeux de données. Nos premières explorations ont suggéré la présence de différences structurales et lexicales – déjà évoquées dans la littérature – entre les phrases sollicitées et artificielles, ainsi que des écarts dans le comportement d'insertion des mots imposés, en particulier pour les mots entretenant une relation dite *syntagmatique* ou *de collocation* (ex : *Sac – Dos*). Les problématiques concernent alors trois axes : (1) caractéristiques générales des phrases (longueur, fréquence et diversité lexicale), (2) caractéristiques inhérentes à la tâche (ordre et flexion des mots imposés), (3) insertion des noms entretenant une relation de collocation (emploi de séquences collocatives).

Ce mémoire s'attache ainsi à observer si les comportements linguistiques des LLMs en tâche de production de phrases avec mots imposés diffèrent de ceux des sujets humains, et dans quelle mesure. Un intérêt particulier est apporté à l'étude du troisième axe qui consiste précisément à questionner l'emploi par les humains et par les LLMs de ce que nous nous proposons d'appeler les *séquences collocatives nominales*. Ces séquences émanent de l'assemblage de deux noms en un syntagme nominal de manière conjointe (ex : « nœud papillon ») ou disjointe (ex : « sac à dos », « clé de sol », « noces d'or », « sucre en poudre »). Ce sont des structures lexicalisées plus ou moins figées du langage naturel. Une partie essentielle de ce travail vise donc à évaluer la capacité des LLMs à proposer ce type de séquences à partir des deux noms qui les composent relativement aux humains, et à expliquer les écarts éventuels.

La méthodologie d'approche de ce travail suit plusieurs étapes successives. Une délimitation précise des objets de l'étude est d'abord nécessaire afin de constituer deux jeux de données équivalents et alignés avec les questions de recherche. Il convient ensuite d'analyser automatiquement les phrases sollicitées et artificielles afin d'en repérer les différences générales (structurales et lexicales), inhérentes à la tâche, et relatives à l'emploi des séquences collocatives. Des analyses statistiques et linguistiques permettent de mettre en avant les contrastes entre les phrases étudiées, et de proposer des explications à ces tendances.

Dans cet écrit, nous commencerons par présenter des notions et concepts en sémantique lexicale et en Traitement Automatique des Langues (TAL), ainsi que des travaux pertinents pour notre étude. Cette étape permettra non seulement de situer le présent travail dans l'état actuel des connaissances et des travaux en linguistique informatique, mais également de guider les questions qui seront traitées et les démarches qui seront effectuées. Nous proposerons ensuite une étude préliminaire des données qui permettra d'expliciter l'émergence des hypothèses de travail. Nous mettrons alors en place le matériel d'étude afin qu'il soit aligné avec nos besoins spécifiques d'observation des séquences collocatives nominales. Nous conduirons une étude comparative des jeux de données selon les trois axes problématisés. Nous étudierons notamment l'emploi des séquences collocatives nominales par les humains et par les LLMs tout en cherchant à expliquer les tendances observées. Enfin, nous conclurons sur les résultats de l'étude avant d'en discuter les apports et les limites et de proposer des perspectives d'études ultérieures.

1 Cadre théorique : relations lexico-sémantiques et LLMs

1.1 Caractérisation des relations entre lexies

La sémantique lexicale étudie le sens des lexies par une description de leurs propriétés sémantiques, ainsi que des relations lexico-sémantiques qui unissent les lexies. Elle définit les différents types de relations qui unissent les unités lexicales à partir de leur forme ou de leur sens lexical respectif. Pour expliciter les définitions ci-dessous, nous nous appuierons au maximum sur des exemples issus des données de notre étude.

1.1.1 Relations lexicales syntagmatiques

Les relations syntagmatiques (situées sur l'axe syntagmatique décrit par De Saussure, 1916) renvoient à la manière dont les lexies se combinent entre elles au sein de la phrase (Polguère, 2003). Elles peuvent ainsi concerner la co-occurrence de lexies dans des collocations telles que « sac à main », « sac à dos », ou dans des expressions idiomatiques telles que « être pris la main dans le sac », « partir sac au dos ». Elles peuvent également désigner les relations de combinaison et de dépendance entre unités lexicales (Polguère, 2003).

Ainsi, deux lexies peuvent entretenir une relation de collocation lorsqu'elles sont associées de manière privilégiée au sein de certaines séquences lexicales, parmi lesquelles figurent notamment ce que Gross (1988, 1996) appelle les *noms composés*. Il s'agit de séquences lexicalisées (c'est-à-dire stabilisées dans les usages) constituées d'un ou plusieurs mots pleins – éventuellement accompagnés de mots grammaticaux – caractérisées par une forte cohésion syntaxique (principe de figement). L'auteur propose une typologie basée sur la structure syntaxique de ces séquences à partir des catégories grammaticales de leurs composants. Il y présente notamment une liste des types de séquences constituées de deux noms : **N N** (ex : *nœud papillon*), **N à N** (ex : *sac à dos*), **N de N** (ex : *sac de sport*), **N en N** (ex : *sac en cuir*), **N par N** (ex : *preuve par neuf*), **N PREP N** (ex **:** *sculpture sur bois*).

Ces séquences se distinguent des expressions idiomatiques, également lexicalisées, dont la signification relève de conventions et ne peut être déterminée par la somme du sens de leurs unités (Marquer, 1994 ; Hattouti et al., 2016) par principe de compositionnalité (Polguère, 2003). Par exemple, en opposition avec la séquence compositionnelle « sac à main » formée à partir des noms *sac* et *main*, l'expression « prendre la main dans le sac », utilisée sous la forme « être pris la main dans le sac » (Díaz, 2009) est non-compositionnelle.

1.1.2 Relations lexicales paradigmatiques

Relations de forme

Situées sur l'axe paradigmatique qui se rapporte aux connexions des lexies à l'intérieur du lexique (De Saussure, 1916), les relations de forme peuvent être graphémiques ou phonétiques (homographie, homonymie). L'homographie est une relation de forme graphémique qui désigne deux mots s'écrivant de manière identique mais n'ayant pas le même sens lexical (ex : le mot *vers* peut se référer à l'unité de ligne poétique, à l'animal invertébré (au pluriel), désigner une direction ou l'approximation d'une heure). L'homophonie, qui est une relation de forme phonétique, s'applique aux mots de formes sonores identiques mais n'ayant pas le même sens, ni la même forme graphémique (ex : *seau*, *saut*, *sot*, *sceau*). Les formes ayant une même racine morphémique peuvent entretenir une relation morphologique de dérivation, qui implique généralement une proximité sémantique (ex : *enseignement*, *enseigner*, *enseignant*).

Relations sémantiques

Les relations sémantiques, se situant également sur l'axe paradigmatique, recensent les relations d'équivalence (synonymie), les relations d'opposition (antonymie), les relations hiérarchiques (hyperonymie, hyponymie et co-hyponymie), les relations partie-tout (méronymie et holonymie), les relations d'instanciation (instance), ou encore l'association de sens (voir 1.1.3 pour les relations lexicales associatives).

La synonymie (du grec συν- (sin) signifiant "avec", "ensemble" et ὄνομα (onoma) "nom"), désigne la relation de similarité entre les signifiés de deux unités lexicales (ex : *trou*, *creux*). Notons que des synonymes absolus (où S1=S2) sont rarissimes puisque les sens de deux mots synonymes seront souvent distingués par quelques nuances. La plupart des synonymes sont ainsi des quasi-synonymes dont le signifié est proche mais pas identique (S1≃S2). Il s'agit de synonymes *approximatifs* (Polguère, 2003), *partiels* ou *contextuels* (Riegel, 1994).

Son contraire, l'antonymie (du grec αντι- (anti) signifiant "au lieu de") renvoie à la relation d'opposition de sens entre les signifiés de deux lexies qui peut être plus ou moins forte et prendre différentes formes (ex : *maladie*, *guérison* désignent des états opposés ; *jour* et *nuit* s'opposent par certaines caractéristiques, mais peuvent être pensés comme des principes complémentaires).

L'hyperonymie (du grec ὑπέρ- (iper) signifiant "haut", "au-dessus") désigne une unité lexicale dont le signifié englobe le sens d'un autre mot dans leur dimension conceptuelle (ex : puisque le signifié du mot *pâte* est une classe conceptuelle qui contient le signifié de *spaghetti*, alors le lexème *pâte* est hyperonyme du mot *spaghetti*).

Sa réciproque, l'hyponymie (du grec ὑπο- (ipo) signifiant "sous") pointe les lexèmes dont le sens est compris par le signifié d'un autre mot (ex : puisque *spaghetti* appartient à la catégorie *pâte*, alors le mot *spaghetti* est un hyponyme du lexème *pâte*).

La co-hyponymie renvoie à deux mots ayant tous deux le même hyperonyme. Ils se situent ainsi sur le même plan hiérarchique (ex : les mots *spaghetti* et *macaroni* sont tous deux inclus dans la classe conceptuelle du lexème *pâte*).

La méronymie (du grec μέρος (méros) signifiant "partie"), est à ne pas confondre avec l'hyponymie. Ici, il ne s'agit pas de classe conceptuelle, mais plutôt de proximité physique des entités observables dans le monde. On peut parler d'une relation de partie-tout dans laquelle le méronyme est la partie. Ainsi, si un référent du monde constitue la partie d'un autre référent (ex : le référent du mot *lave* est une partie de son tout qu'est le référent du mot *volcan*), alors son signifiant sera méronyme du mot de son tout (dans notre exemple, *lave* est donc méronyme de *volcan*).

Sa réciproque, l'holonymie (du grec ὅλον (olon) signifiant "tout") désigne le tout (ex : le mot *volcan* est holonyme du mot *lave*).

L'instance, quant à elle, renvoie à une relation dans laquelle un lexème catégoriel est associé à une expression référentielle unique – souvent sous forme de nom propre – désignant un individu, un objet, un lieu, ou une entité précise du monde, appartenant à la catégorie conceptuelle exprimée par le premier (ex : *Toulouse* est une instance de *ville*).

1.1.3 Relations lexicales associatives

Les lexies peuvent être liées par une relation lexicale associative également appelée *association d'idées*. Il s'agit d'une relation sémantique dite "non-classique" par opposition aux relations sémantiques décrites précédemment. La relation associative désigne la relation entre deux mots dont on ne peut observer comme lien sémantique qu'une association d'idées construite selon des représentations culturelles du monde, ou en discours (ex : *maladie* et *hôpital*). Les relations lexicales associatives sont, par ailleurs, plus fréquentes en contexte que les relations sémantiques (Morris & Hirst, 2004).

Morris et Hirst (2004) précisent que les mots entretenant une relation lexico-sémantique dite "classique" (Cf. relations présentées en partie 1.1.2) relèvent systématiquement de la même catégorie grammaticale, et partagent certaines propriétés définitoires. Les mots en relation associative ne partagent pas nécessairement la même catégorie grammaticale (ex : *soleil* et *chaud*) et ne possèdent jamais les mêmes traits définitoires (ex : *parc* et *enfant*). Le lien entre ces lexies est souvent perçu comme évident auprès des locuteurs bien qu'il ne soit ni hiérarchique (pas d'organisation d'imbrication des lexies), ni dépendant d'une classe grammaticale (les relations associatives peuvent relier des lexèmes de toute nature entre eux). Ces associations sont d'ordre conceptuel, culturel et cognitif, et très bien ancrées chez les locuteurs.

Morris et Hirst (2004) soutiennent que les systèmes de Traitement Automatique du Langage (TAL) ne peuvent restituer adéquatement la perception humaine de la cohésion lexicale s'ils se limitent aux seules relations sémantiques classiques. Il est donc nécessaire d'intégrer l'identification des liens sémantiques non-classiques qui unissent les mots dans le traitement automatique des textes.

1.1.4 Relations lexicales multiples

Il est possible que deux lexèmes se voient attribuer plusieurs relations sémantiques (ex : *spaghetti* et *bolognaise* peuvent constituer une collocation lorsque « spaghettis bolognaise » s'emploie comme séquence figée désignant un référent spécifique (plat/recette). Leur lien sémantique peut également relever d'une association d'idées reposant sur l'association des concepts plus généraux des mots *spaghetti* (pâtes) et *bolognaise* (sauce) dans le lexique mental des locuteurs. Le placement des unités lexicales concernées dans des phrases permet ainsi d'apporter un contexte linguistique, de guider l'analyse sémantique et de mieux comprendre les concepts mobilisés par le locuteur et, ainsi, de mieux catégoriser le type de relation sémantique entre les mots cibles dans le contexte donné. Pour la paire de mots *Spaghetti – Bolognaise*, les relations lexicales multiples sont donc déliées comme ci-après : la collocation est marquée dans des énoncés tels que « Ce midi, on mange des spaghettis bolognaise. » et l'association d'idées est mise en avant dans des phrases comme « Je préfère mettre de la bolognaise dans mes spaghettis. ».

1.1.5 Marquage des relations lexicales classiques dans le discours

Lorsque deux mots entretenant une relation lexicale classique se retrouvent insérés en discours, certains motifs peuvent émerger dans les phrases. Taher & Salik (2022) précisent que des énoncés comprenant deux mots en relation de méronymie peuvent être marqués par des motifs tels que "Un X est une partie d'un Y" ou "Un Y a un/des X". Pour exemples : « Un doigt est une partie de la main », « Une main a des doigts ». Aussi, les phrases contenant deux mots en relation d'hyperonymie peuvent suivre des motifs tels que "Le X est un Y", "Les Y, et notamment les X" ou "Les Y, dont les X". Pour exemples : « Le cœur est un organe », « Les organes, et

notamment le cœur […] », « Les organes, dont le cœur, […] ». En relation d'hyperonymie, l'exception peut également être marquée dans le discours par des énoncés du type "Les Y, sauf X" ou "Les Y, à l'exception de X", l'inclusion par des phrases telles que "Les Y, y compris X", et l'exemplification par des énoncés tels que "Les Y comme X" ou "Les Y tels que X" (Cruse, 1986). On peut ainsi supposer, en s'inspirant de ces patrons, que le marquage de la co-hyponymie en discours pourrait suivre un motif tel que "Les X_1 et les X_2 sont des Y", "Les Y, et notamment les X_1 et les X_2" ou "Les Y, dont le X_1 et le X_2". Pour exemples : « Les chiens et les chats sont des animaux », « Les *animaux*, et notamment les chiens et les chats […] », « Les animaux, dont les chiens et les chats […] ».

Ces motifs ne couvrent cependant qu'une partie des modes d'association de deux mots en contexte. Nous nous attendons à ce que les phrases issues du discours soient plus variées dans le marquage des relations lexicales. On peut donc imaginer des phrases suivant des motifs alternatifs, comme dans les exemples suivants, issus d'une tâche de production de phrases avec mots imposés en relation de méronymie : « Il y a de la lave dans le volcan. » ou « Le biberon contient du lait. ». Aussi, ces motifs sont souvent très ambigus, comme « Le X est un Y » qui peut correspondre à d'autres cas que l'hyperonymie (ex : « Le chat est une nuisance. »).

Notons que les lexies synonymiques se retrouvent rarement en co-occurrence dans des énoncés naturels et ont plutôt tendance à se substituer dans des contextes linguistiques identiques (Rapp, 2002). Les motifs associés à la synonymie relèvent davantage de reformulations du type « X, également appelé Y » ou bien « X, autrement dit Y ». On peut également les retrouver dans des cas de coordination (ex : « Pierres et cailloux jonchaient le sol. »). Nous nous attendons à ce que deux mots synonymes en co-occurrence soient employés de façon métalinguistique, c'est-à-dire pour renvoyer à eux-mêmes en tant qu'unités de la langue (ex : « Tu dis *un sac* ou *une poche* ? »). Si les lexies insérées en co-occurrence sont des quasi-synonymes, nous supposons qu'on y repérera une mise en évidence des nuances sémantiques qui les distinguent à travers le contexte linguistique (ex : « Un sachet est un petit sac. »).

1.1.6 Identification automatique des relations lexico-sémantiques

La tâche d'association lexicale consiste à présenter un mot stimulus à un participant et à lui demander de produire le premier mot qui lui vient à l'esprit. Les réponses obtenues à cette tâche peuvent refléter une grande diversité de relations lexico-sémantiques entre le mot stimulus et le mot répondu. Dans le cadre du projet *Evolex* – dont l'objectif est d'étudier la fluence et l'accès lexical, Gaume et al. (2018) ont analysé manuellement 559 paires de mots stimulus-réponse pour une tâche d'association lexicale basée sur 60 stimuli nominaux soumis à 30 participants. Leurs annotations montrent que les relations sémantiques classiques représentent près de la moitié des paires stimulus-réponse analysées (49,5%), les relations associatives non-classiques en constituent plus d'un tiers (36,1%) tandis que les relations syntagmatiques se retrouvent plutôt sous-représentées (8,8%) avec les relations de formes phonologiques (0,9%).

Gaume et al. (2018) ont ainsi cherché à savoir si les méthodes automatiques de traitement du langage pouvaient identifier adéquatement le type de relation lexicale entre un stimulus et sa réponse. Un corpus général et un dictionnaire ont été utilisés pour mesurer la similarité entre les mots stimuli et les réponses des participants à l'aide de mesures de similarité sémantique. Les résultats ont montré que si ces méthodes automatiques s'avèrent performantes pour repérer les relations sémantiques classiques, elles restent peu efficaces pour identifier les relations associatives.

1.2 LLMs : traitement et génération textuelle

Ancrés dans le domaine du TAL, les modèles de langage (LMs) sont des systèmes qui estiment, à partir des régularités apprises en corpus, la suite la plus probable d'un texte. Plus précisément, les LMs calculent la probabilité qu'un mot apparaisse après une séquence donnée (Jurafsky & Martin, 2025). Les modèles *n-grams* sont les LMs qui possèdent l'architecture la plus simple. Ces derniers calculent la probabilité d'apparition d'un mot à partir des *n*–1 mots précédents (ex : un modèle de trigramme utilise une fenêtre *n*=3 et calcule la probabilité du mot suivant *n*–1 à partir d'une séquence de deux mots).

Dans le domaine de l'Intelligence Artificielle (IA), les *Large Language Models* (LLMs) sont une extension à grande échelle des modèles de langage traditionnels. Ils sont conçus pour traiter une séquence textuelle et générer un texte cohérent par calculs de probabilités grâce à leur apprentissage sur de larges jeux de données et à un nombre massif de paramètres.

1.2.1 LLMs : fonctionnement théorique

LLMs : tâches, *fine-tuning*

Yang et al. (2024) décrivent la capacité des LLMs à générer la suite probable d'un texte à partir de calculs statistiques, une propriété fondamentale qui leur permet d'accomplir de multiples tâches linguistiques. On distingue alors l'utilisation de modèles généralistes de grande taille (LLMs) capables de s'adapter à un large éventail de tâches de natures diverses, et le recours à des modèles *fine-tuned* spécialisés dans une tâche ou un type de texte donné. La technique de *fine-tuning* consiste à ajuster un modèle pré-entraîné sur un jeu de données restreint afin de le spécialiser. Les modèles *fine-tuned* sont performants lorsque la tâche est clairement définie. Cependant, leurs performances chutent lorsque les données ou la tâche s'éloignent de leur domaine d'entraînement. A l'inverse, les LLMs sont plus robustes sur des tâches ouvertes et peu formalisées.

Tâches de génération

Yang et al. (2024) distinguent deux grandes catégories de tâches de génération : la conversion de texte d'entrée qui consiste en la génération de nouvelles séquences à partir du texte d'entrée (synthèse de texte, traduction automatique) et la génération libre (*open-ended generation*) qui concerne la génération de texte à partir d'une simple description de la tâche ou d'instructions (rédaction d'e-mails, composition d'articles de journal, création d'histoires fictives, écriture de code informatique).

Jeux de données pour le pré-entraînement

Les jeux de données employés pour le pré-entraînement des LLMs sont généralement constitués exclusivement de ressources textuelles : livres, articles, pages web. La qualité, quantité et diversité de données d'entraînement influencent significativement la performance des LLMs. La nature de ces données agissent également sur les capacités des LLMs et les conditionne dans les tâches qu'ils effectueront. Par exemple, certains modèles ont été entraînés sur des jeux de données avec une plus large quantité de données multilingues, et auront ainsi de meilleures performances dans des tâches multilingues telles que la traduction, sans même avoir été explicitement entraînés pour ces tâches (Yang et al., 2024).

Paramètres des LLMs et performance

Yang et al. (2024) précisent que la performance des LLMs augmente avec le nombre de paramètres qui les composent, c'est-à-dire le poids des connexions du réseau de neurones artificiels des modèles. Il convient de noter que les plus grands LLMs actuels comptent des centaines de milliards de paramètres.

Trois grands types de LLMs

Yang et al. (2024) cernent trois grands types de LLMs qui diffèrent selon la stratégie employée pour leur entraînement, leur modèle d'architecture et leur type d'utilisation : il s'agit des modèles à attention causale (*causal attention models*), des modèles à attention non-causale (*non-causal attention models*), et des modèles à attention semi-causale (*half-causal attention models*).

Les modèles à attention causale, aussi appelés auto-régressifs (*autoregressive models*) ou *decoder-only*, sont *fine-tuned* pour être plus performants sur des tâches cibles : généralement la génération textuelle ou la génération de réponses aux questions. Ces types de modèles lisent le texte en entrée *token* par *token* et prennent ainsi en compte l'ordre syntaxique de l'énoncé fourni, comme c'est le cas lors d'un échange en langage naturel. Parmi les modèles auto-régressifs, nous pouvons citer : GPT-3, Claude, LlaMA, Bard, OPT, BLOOM.

Les modèles à attention non-causale, également appelés *encoder-only*, utilisent le *masked language modeling*, soit la prédiction d'un mot masqué grâce à son contexte linguistique gauche et droit. Contrairement au modèle auto-régressif, le modèle à attention non-causale voit simultanément tous les *tokens* du texte fourni en entrée. Ce type de modèles est décrit comme particulièrement performant en tâche de compréhension du langage. Des exemples de modèles à attention non-causale sont : BERT, ALBERT, RoBERTa, DeBERTa.

Les modèles à attention semi-causale, *encoder-decoder*, combinent un traitement du texte avec attention non-causale, c'est-à-dire une vision simultanée de tous les *tokens* en entrée, et une génération textuelle par attention causale, soit une génération séquentielle, *token* par *token*. Deux exemples de modèles à attention semi-causale sont : Flan-UL2 et Alexa TM.

Parmi ces trois types de modèles, les *decoder-only* constituent aujourd'hui le type dominant de LLMs. Ce type de modèles est devenu populaire avec l'arrivée de GPT-3 sur le marché de l'IA en 2021.

Modèles *open-source* et *closed-source*

Durant les premières années suivant l'apparition des agents conversationnels (systèmes conçus pour simuler une interaction en langage naturel avec un utilisateur) auprès du public, les LLMs recensaient une majorité de modèles *open-source*, dont le code est visible et modifiable par tous dans le but que chacun puisse contribuer à l'amélioration de la ressource. Depuis l'arrivée de GPT-3, modèle *closed-source*, la plupart des nouveaux LLMs publiés sont également *closed-source*. Leur code n'est donc accessible que par l'organisation propriétaire du modèle. Cela permet aux organisations de maintenir leur concurrence sur le marché de l'IA, mais rend plus difficiles les études sur l'architecture et l'entraînement des LLMs par les chercheurs (Yang et al., 2024).

Modèles *open-weight* et *closed-weight*

Aussi, les développeurs peuvent publient les paramètres de leur modèle afin qu'il soit téléchargé et exécuté par des tiers sans donner accès aux données d'entraînement exactes ou au code source : il s'agit des modèles *open-weight*. Le modèle et ses paramètres peuvent rester strictement confidentiels et hébergés par les développeurs. Dans ce cas, l'accès au LLM se fait exclusivement via une interface en ligne : ce sont les modèles *closed-weight* (Solaiman, 2023).

Limites des LLMs

Yang et al. (2024) attirent l'attention sur les limites des LLMs et les biais existant dans leurs réponses. En effet, les LLMs peuvent produire des réponses plausibles mais erronées par rapport aux connaissances humaines et aux faits réels et connus : ces réponses sont appelées *hallucinations*. Les auteurs précisent également que les LLMs sont très sensibles aux instructions reçues en entrée (voir 1.2.2 sur le *prompting*).

1.2.2 Prompting

La qualité des sorties des LLMs est influencée par de nombreux paramètres concernant son architecture, mais également par les données qu'il reçoit en entrée. Dans leur article, Schulhoff et al. (2024) définissent le terme de *prompt* comme une entrée (*input*) fournie à un modèle d'IA générative avec pour but de guider la sortie (*output*) selon des résultats attendus, et/ou selon un format désiré par l'utilisateur. Un prompt contient généralement une instruction textuelle et peut comporter des fichiers supplémentaires comme des documents textuels, des images, des audios, et autres types de fichiers. Le terme *prompting* désigne l'ensemble des techniques utilisées pour formuler les instructions adressées à un modèle afin d'améliorer la qualité de ses réponses.

Afin de guider précisément le format désiré en sortie, l'utilisateur peut faire usage d'exemples au sein de son prompt. Il s'agit de spécifier le format attendu au travers d'exemples. Cette technique d'exemplification au sein du prompt est nommée *In-Context Learning* (ICL). Selon la tâche considérée, les exemples fournis dans un prompt peuvent amener le modèle à obtenir de meilleures performances. Il existe plusieurs types de méthodes *shot-based* (applications directes de l'ICL) où *shot* désigne la quantité d'exemples transmis au modèle pour une tâche donnée. Lorsque le modèle apprend à compléter la tâche à l'aide de quelques exemples fournis, il s'agit de *few-shot prompting*. Lorsqu'un seul exemple est dispensé, on parle de *one-shot prompting*. Enfin, lorsqu'aucun exemple n'est indiqué, il s'agit de *zero-shot prompting* (Schulhoff et al., 2024).

1.2.3 Génération textuelle et cohérence sémantique

Les LLMs sont aujourd'hui capables de générer des histoires sémantiquement cohérentes et semblables à ce qu'un humain pourrait rédiger (Sanacore, 2024 ; Muñoz-Ortiz et al., 2024). Sanacore (2024) montre que ChatGPT génère des histoires mobilisant des scénarios plausibles et des associations lexicales proches de celles observées dans les productions humaines. Selon l'auteur, cette cohérence repose notamment sur la capacité du modèle à reconnaître des proximités sémantiques entre les mots et à réutiliser des concepts prototypiques appris lors de son entraînement.

Les travaux de Cai et al. (2024) ont confirmé ces capacités dans une tâche plus ciblée : la génération de phrases-exemples d'usage d'un mot dans un dictionnaire. L'objectif de leurs démarches était de fournir une mesure des capacités des modèles à exemplifier et illustrer la signification des mots. Leurs résultats ont montré que ces productions étaient fluides,

grammaticalement correctes et proches des exemplifications rédigées par des humains. Ils ont également noté que la qualité était meilleure pour l'exemplification de mots fréquents que pour les mots rares ou spécialisés.

Néanmoins, ces performances sémantiques ne doivent pas être confondues avec des capacités de compréhension du langage humain. En effet, les modèles ne "comprennent" pas la tâche qui leur est assignée au travers des instructions textuelles : ils repèrent des patterns linguistiques de surface grâce à leur entraînement sur d'immenses corpus textuels, et mobilisent les régularités statistiques apprises pour prédire la suite statistique logique du texte fourni en entrée par l'utilisateur.

1.2.4 Identification différentielle de textes produits par l'humain ou par l'IA

La proximité entre les textes produits par les humains et ceux générés par les LLMs soulève plusieurs questions : est-il possible d'identifier clairement l'origine d'un texte et son auteur ? Les textes que nous lisons au quotidien ont-ils été rédigés par des humains, ou générés par l'IA ? Cette problématique est devenue centrale dans le domaine de l'IA et auprès de la population.

Alavoine et al. (2024) ont conduit une étude dans laquelle ils ont cherché à savoir comment les humains percevaient les textes générés par l'IA par rapport à des textes rédigés par l'humain. Leur jeu de données recensait des questions ainsi que des textes produits en réponse aux questions par des sujets humains sur un forum en ligne, et des réponses générées par ChatGPT à partir des mêmes questions. Les chercheurs ont conduit une campagne d'annotation sur ce jeu de données : 17 participants ont annoté les réponses qui leur étaient présentées en statuant s'il s'agissait d'un texte rédigé par un humain ou par l'IA. Les résultats montrent que, dans la majorité des cas (81%), les participants étaient capables de distinguer les réponses artificielles des réponses humaines. Les chercheurs ont également précisé que la tâche d'annotation était plus aisée pour les participants lorsque le texte à classer était plus court.

En complément de cette expérience, Alavoine et al. (2024) ont comparé les caractéristiques syntaxiques et lexicales des deux types de réponses (humaines et artificielles) issues du même jeu de données pour en noter les différences linguistiques. Au niveau syntaxique, ils ont notamment pu observer un écart de longueur entre les deux types de réponses. Selon eux, les humains répondent de façon plus concise aux questions qui leur sont posées (30% des réponses comptent moins de 500 caractères) tandis que ChatGPT a tendance à générer des énoncés plus longs mais d'une longueur beaucoup plus uniforme (entre 1000 et 1500 caractères). Au niveau lexical, les chercheurs ont pu observer – grâce à un calcul de *Type-Token Ratio* (TTR) (nombre de mots différents sur le nombre total de mots d'un texte) – une richesse lexicale plus étendue dans les réponses des sujets humains que dans celles générées par ChatGPT.

Malgré les capacités des LLMs à générer du texte proche du langage naturel, certains traits syntaxiques et sémantiques semblent tout de même permettre aux lecteurs d'identifier correctement l'origine d'un texte de manière intuitive.

1.2.5 Analyse contrastive des productions textuelles d'humains et de LLMs

Aussi, dans leur étude, Muñoz-Ortiz et al. (2024) ont comparé les caractéristiques linguistiques de textes d'actualité en anglais produits par des sujets humains et par des LLMs. Pour ce faire, les auteurs ont mis en parallèle un corpus journalistique et des textes générés par six LLMs différents. Ils ont ainsi comparé ces textes afin d'en étudier les contrastes lexicaux, morphosyntaxiques et sémantiques.

Pour mesurer la diversité lexicale, les chercheurs ont employé le *Standardized Type-Token Ratio* (STTR) ainsi que la *Measure of Textual Lexical Diversity* (MTLD). L'analyse morphosyntaxique a été automatisée avec l'outil *Stanza* pour la segmentation, la tokenisation, l'étiquetage morphosyntaxique, l'analyse syntaxique des dépendants et des constituants. La similarité sémantique entre les textes a été mesurée grâce à un modèle d'*embeddings*. Ce dernier permet de représenter les mots sous forme de vecteurs et de mettre en avant les concepts sémantiquement proches grâce à une cartographie de proximité des vecteurs. Des tests statistiques *t-tests* ont permis aux chercheurs de vérifier la significativité des métriques employées dans leur étude.

D'après leurs résultats, les textes de LLMs présentent certaines différences avec les textes des sujets humains (confirmant les conclusions d'Alavoine et al., 2024). Le vocabulaire employé par les LLMs est plus restreint. Les modèles de la famille LLaMA figurent parmi les LLMs dont la richesse lexicale se rapproche le plus de celle des humains. Les productions des LLMs montrent une préférence pour les syntagmes verbaux. Dans les textes des sujets humains, les constituants syntaxiques sont plus courts, les phrases sont plus longues, appelant un usage plus répété de la ponctuation. Les sujets humains préfèrent la formulation de phrases nominales tandis que les textes des LLMs reposent sur des structures principalement verbales. A noter que, chez Alavoine et al. (2024), ce sont les humains qui produisaient les textes les plus courts. On peut supposer que cette différence est due au contexte de production des énoncés : réponses à des questions sur un forum tandis que Muñoz-Ortiz et al. (2024) s'appuyaient sur des textes d'actualité. Au niveau sémantique, les chercheurs ont observé une grande similitude textuelle avec les textes de sujets humains. La taille des LLMs n'entraîne pas de différence majeure de similarité entre les deux types de productions. Les écarts de productions entre les LLMs eux-mêmes sont également négligeables.

1.2.6 Classification des associations de mots par les LLMs

Outre leur capacité à générer des textes cohérents et proches du langage naturel, les recherches en TAL s'intéressent également à la capacité des LLMs à identifier les relations lexico-sémantiques entre les unités lexicales (pour les relations lexico-sémantiques, Cf. 1.1).

Dans le cadre de leur recherche, Rodriguez et al. (2025) ont mesuré les performances de trois modèles de langage de tailles différentes (GPT-4o-mini (petit), Llama3.1-70b (moyen) et GPT4o (grand)) sur la classification de mots issus d'une tâche d'association lexicale par leur relation lexico-sémantique. Après avoir proposé une taxonomie des relations lexico-sémantiques, les auteurs ont appliqué leur taxonomie à un jeu de données contenant 1300 associations de mots. Ils ont ensuite fait classifier ces associations de mots selon leur taxonomie aux différents modèles afin de mesurer leurs capacités de classification. Les auteurs expliquent que les modèles comme GPT-4o obtiennent de faibles performances dans cette tâche, questionnant leurs capacités à traiter et remployer les principes sous-jacents de l'association de mots chez l'humain.

1.2.7 Détection des expressions idiomatiques par les LLMs

Aussi, certains auteurs se sont intéressés à l'identification des relations syntagmatiques (Cf. 1.1.1) par les LLMs. Parmi eux, Hashiloni et al. (2025) ont observé la détection des *multi-word expressions* (*MWEs*) – séquences de lexies non-compositionnelles, c'est-à-dire dont le sens global n'est pas interprétable par la somme des sens de leurs unités – par les LLMs. Les MWEs peuvent être des expressions idiomatiques, des noms composés ou des collocations et sont inhérentes à chaque langue. Les auteurs se sont plus précisément penchés sur une sous-catégorie des MWEs : les expressions idiomatiques. L'identification des expressions idiomatiques est centrale en TAL – notamment pour des tâches telles que la traduction automatique ou la représentation sémantique – puisqu'une interprétation erronée de ces séquences pourrait nuire aux performances des outils. Par exemple, l'interprétation littérale de la séquence « nœud papillon » pourrait mener à la traduction mot à mot « knot butterfly » ou « bow butterfly » au lieu de « bow tie ». La méthode des chercheurs consistait à fournir des phrases issues de plusieurs jeux de données (dont un multilingue) au LLM (non *fine-tuned*) en plus d'instructions lui demandant d'identifier les expressions idiomatiques dans les segments proposés. Les LLMs évalués étaient : GPT-4o, GPT-4o-mini, Llama-4-Scout, Qwen2.5, et deux modèles de raisonnement (GPT-o3-mini, DeepSeek-R1). Les chercheurs ont utilisé différents types de prompts en *zero-shot* ou en *few-shot*. Ils ont ensuite comparé les résultats des modèles évalués avec ceux de modèles supervisés entraînés sur des données annotées et des modèles *fine-tuned* entraînés sur cette tâche spécifique. A partir de leurs résultats, les auteurs ont conclu que certaines techniques de *prompting* pouvaient guider efficacement les LLMs vers une réflexion linguistique aidant ces derniers à identifier adéquatement les expressions idiomatiques. Cependant, en élargissant l'objet d'identification aux MWEs, les chercheurs ont noté que l'identification de ces unités structurellement variées et de nature flexible et inconstante pouvait présenter un challenge pour les LLMs.

Cette revue de la littérature a permis de présenter les principales relations lexico-sémantiques unissant les lexies ainsi que différentes approches pour les caractériser et les identifier. Elle a également mis en évidence les capacités des LLMs à traiter des phénomènes linguistiques variés, mais aussi leurs limites, notamment dans l'identification de certaines séquences lexicalisées du langage naturel, telles que les *multi-word expressions*. Par ailleurs, cette revue a souligné l'intérêt de la littérature pour l'analyse contrastive des textes produits par l'humain et par les LLMs. Dans la continuité de ces travaux, notre étude vise à comparer des phrases produites par des humains et des phrases générées par des LLMs pour une même tâche. L'objectif est d'en repérer certaines différences linguistiques, mais aussi d'analyser précisément la capacité des LLMs à produire des séquences lexicalisées plus ou moins figées du langage naturel par rapport aux humains.

2 Étude préliminaire des données

Dans cette partie, nous présenterons d'abord nos deux jeux de données (phrases sollicitées et phrases artificielles), puis nous décrirons l'étude préliminaire conduite sur deux échantillons de test qui nous a amené à formuler nos hypothèses de travail.

2.1 Phrases sollicitées : matériel préalable

2.1.1 Evolex : paires de mots

Une première expérience a été effectuée dans le cadre du projet *Evolex* – dont l'objectif premier concerne l'étude de la fluence et de l'accès lexical. L'expérience Evolex proposait une tâche d'association lexicale avec amorçage. Les étapes de cette première expérience Evolex sont illustrées en orange sur la figure 1. Il s'agissait de placer des participants humains en conditions d'expérimentation, et de leur présenter un nom lu par synthèse vocale, constituant le stimulus (ex : *sac*). Ensuite, il leur était demandé de prononcer un nom qui leur venait en tête suite à l'écoute du stimulus (ex : *dos*). Afin de standardiser ces tests, les consignes étaient identiques pour tous les participants, et les stimuli étaient tous des noms communs de 1 à 3 syllabes au singulier. 30 participants ont écouté 60 stimuli, donnant 1800 réponses au total. Après un tri des données visant à ne conserver que les noms sous leur forme canonique, 1544 réponses ont été conservées, correspondant à 559 paires de mots distinctes (Gaume et al., 2018). Chaque réponse retenue a été regroupée avec le stimulus auquel elle était associée pour former des paires de mots (ex : *Sac – Dos*). Des juges experts ont annoté la relation lexico-sémantique qui unit les deux mots de chaque paire parmi la liste suivante définie par les annotateurs : collocation, synonymie, antonymie, hyperonymie/hyponymie, co-hyponymie, méronymie/holonymie, instance, association de sens (ex : la paire *Parc – Attraction* annotée comme collocation). A noter que les paires étaient annotées en dehors de tout contexte linguistique. Ainsi, des cas d'ambiguïté pouvaient amener les juges experts à annoter plusieurs relations pour une même paire de mots (ex : la paire *Spaghetti – Bolognaise* a été annotée comme collocation ou association de sens).

2.1.2 Evolex : questionnaires et phrases produites

Les paires de mots issues de cette première expérience Evolex ont été reprises et utilisées dans le cadre de questionnaires en ligne créés par des étudiants de Sciences du langage sous la direction de leurs enseignants. Les étapes relatives aux questionnaires Evolex sont illustrées en bleu sur la figure 1. Dans ces derniers, différentes paires de mots Evolex étaient présentées à des participants humains hors contexte linguistique. Il leur était demandé de sélectionner la relation lexico-sémantique qu'ils percevaient entre les deux mots de chaque paire parmi la liste définie par les juges experts. Ensuite, les participants étaient soumis à une tâche de production de phrases avec mots imposés. Ils devaient ainsi rédiger une phrase contenant les mots de la paire, pour chacune des paires présentées, sans contrainte d'ordre d'agencement ou de nombre grammatical des mots imposés. Par exemple, pour la paire de mots *Sac – Dos*, les participants pouvaient rédiger une phrase telle que « Regarde ce mec de dos avec son sac dora ! ». Enfin, il leur était demandé de situer la proximité sémantique entre les mots des paires sur une échelle de 0 à 5.

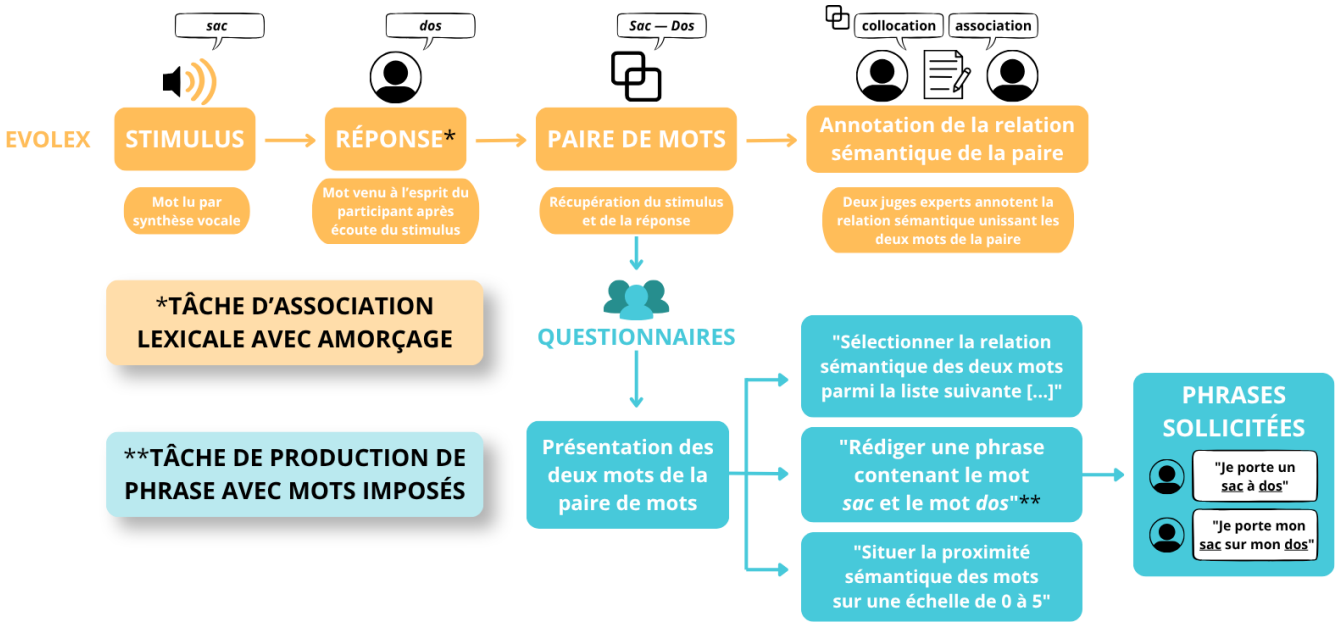

Figure 1 : Protocole de récolte des données Evolex et leur utilisation dans les questionnaires (inspiré de Simounet, 2021)

Nous avons récupéré l'entièreté des phrases des questionnaires à notre disposition (4387 phrases pour 159 paires distinctes).

2.1.3 Formation de l'échantillon de test de phrases sollicitées

Notre objectif était de comparer des phrases sollicitées et artificielles produites à partir des mêmes paires de mots. Nous avons d'abord réduit le matériel préalable (à l'origine 4387 phrases pour 159 paires) afin d'obtenir un échantillon équivalent en termes de représentation des relations lexico-sémantiques des paires de mots. Pour ce faire, nous nous sommes basé sur les annotations des relations lexicales effectuées par les juges experts dans le cadre du projet Evolex. Nous avons conservé 10 paires de mots par relation lexico-sémantique sans critère particulier. Notre jeu de paires de mots présentait cependant une sous-représentation des relations de synonymie (7 paires de mots) et d'antonymie (1 paire de mots) due au manque de réponses lors de la première expérience Evolex. Les paires sélectionnées sont présentées selon la catégorie de relation lexico-sémantique dans l'ordre [Stimulus – Réponse] dans les tableaux 1-A et 1-B.

PAIRES Collocation	PAIRES Antonymie	PAIRES Synonymie
Noce – Nuit	Maladie – Guérison	Caillou – Pierre
Noce – Or	–	Marionnette – Pantin
Parc – Attraction	–	Noce – Mariage
Parc – Huître	–	Sac – Poche
Sac – Bandoulière	–	Sac – Sachet
Sac – Dos	–	Trou – Creux
Sac – Main	–	Trou – Orifice
Sac – Plastique	–	–
Spaghetti – Bolognaise	–	–
Trou – Souris	–	–

Tableau 1-A : Paires de mots sélectionnées pour constituer l'échantillon de test

PAIRES Hyperonymie/ Hyponymie	PAIRES : Co-hyponymie	PAIRES Méronymie/ Holonymie	PAIRES Association
Entrecôte – Viande	Caillou – Rocher	Biberon – Tétine	Biberon – Bébé
Hirondelle – Oiseau	Entrecôte – Steak	Caillou – Montagne	Maladie – Hôpital
Maladie – Rhume	Maladie – Infection	Entrecôte – Bœuf	Marionnette – Enfant
Maladie – Virus	Ortie – Menthe	Marionnette – Fil	Noce – Bague
Marionnette – Jouet	Ortie – Pissenlit	Ortie – Soupe	Ortie – Démangeaison
Ortie – Plante	Parc – Jardin	Osier – Chaise	Parc – Ballade
Osier – Matière	Sac – Cartable	Osier – Panier	Rail – Train
Sac – Bagage	Sœur – Frère	Sac – Anse	Sac – Cours
Spaghetti – Pâte	Sœur – Mère	Sac – Crocodile	Spaghetti – Fourchette
Terrier – Trou	Trou – Vide	Volcan – Lave	Trou – Chute

Tableau 1-B : Paires de mots sélectionnées pour constituer l'échantillon de test

Notons que plusieurs paires partagent un mot commun ici : il s'agit du mot-stimulus de l'expérience Evolex (sauf pour la paire *Terrier – Trou* pour laquelle le mot *Trou* ne constituait pas le stimulus contrairement aux paires *Trou – Chute*, *Trou – Creux*, *Trou – Orifice* et *Trou – Vide*). Les paires conservées recensent 17 stimuli différents.

Nous avons ainsi récupéré toutes les phrases disponibles pour chaque paire de mots conservée, nous donnant un échantillon de 1659 phrases sollicitées à partir des 58 paires de mots.

2.2 Phrases artificielles : échantillon de test

Notre second jeu de données, créé exclusivement pour ce travail de recherche, regroupe des phrases générées par un LLM. Les étapes de génération de ces phrases sont illustrées en figure 2. La consigne donnée à l'IA est similaire à celle donnée aux participants des questionnaires, mais se concentre uniquement sur la génération de phrases avec mots imposés d'après les paires de mots Evolex sans identification préalable de la relation lexico-sémantique.

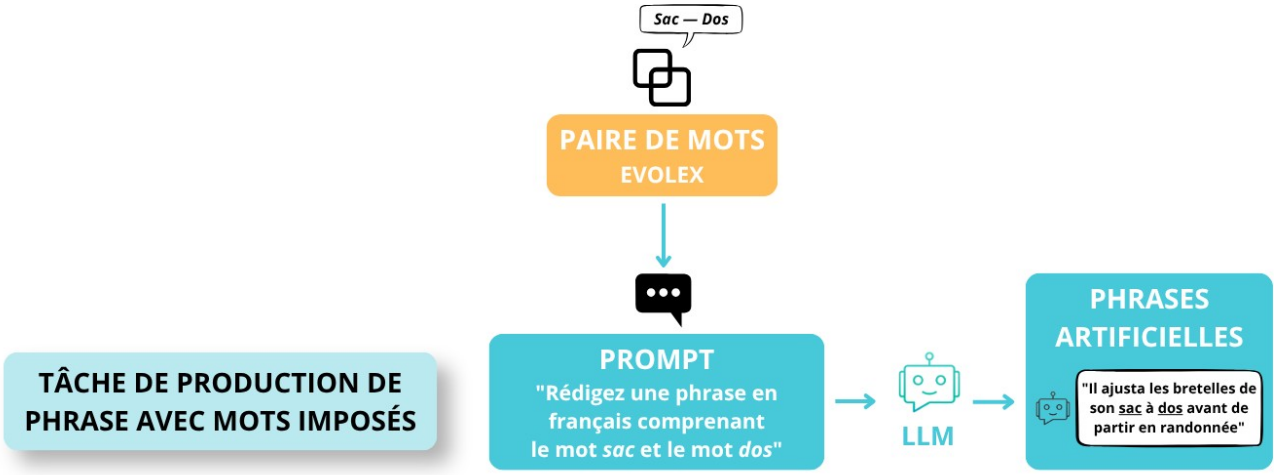

Figure 2 : Protocole de récolte des données artificielles à partir des données Evolex (inspiré de Simounet, 2021)

Pour des raisons d'accessibilité, nous avons sélectionné GPT-5.2 (OpenAI, 2025a) et utilisé son interface en ligne. Nous avons défini un prompt générique à fournir au modèle : « Rédigez une phrase en français comprenant le mot "…" et le mot "…" ». Pour chaque requête, les instructions comportaient ainsi les deux mots d'une paire de mots Evolex dans l'ordre [Stimulus – Réponse]. Les noms étaient tous présentés sous leur forme canonique, au singulier. Aucune instruction n'était spécifiée au modèle concernant le nombre grammatical des noms à employer. L'ordre des mots à adopter n'était pas contraint et aucune mention du lien lexical qui unit les mots de la paire n'était apportée. Pour des besoins d'équivalence entre les deux échantillons à comparer (phrases sollicitées et phrases artificielles), nous avons fourni au modèle les 58 mêmes paires de mots que pour l'échantillon des 1659 phrases sollicitées. Nous avons alors fait générer 20 phrases par paire au modèle, donnant un échantillon de 1160 phrases générées par GPT-5.2 à partir des 58 paires de mots. A noter que les prompts étaient fournis en sessions individuelles et sans accès identifié, évitant ainsi que le modèle ne prenne en compte le contexte précédent pour la génération des nouvelles phrases.

Cet échantillon de test nous a ainsi permis de conduire des premières observations sur les productions d'un LLM en tâche de génération de phrases avec mots imposés en les comparant avec des phrases produites par des sujets humains afin de faire émerger des hypothèses.

2.3 Premières observations

Nous avons conduit des premières observations des deux échantillons constitués (1659 phrases sollicitées et 1160 phrases générées par GPT-5.2) à l'aide de différents outils automatisés par programme Python. L'objectif était de vérifier certaines différences linguistiques entre les productions des humains et des LLMs évoquées dans la littérature, mais également d'explorer d'autres caractéristiques.

Vérifications de la littérature (Alavoine et al., 2024 ; Muñoz-Ortiz et al., 2024)

La mesure du nombre de mots moyen par production (délimités par frontières de mot autour de caractères alphanumériques grâce à l'expression régulière "\b\w+\b") nous a permis de noter un écart de longueur entre les phrases sollicitées et les phrases artificielles : sur nos échantillons, nous relevons une moyenne de 9,77 mots par production chez les sujets humains, contre 8,20 mots par production chez GPT-5.2.

Aussi, nous avons utilisé la *Measure of Textual Lexical Diversity* (MTLD) (Muñoz-Ortiz et al., 2024) sur la totalité des tokens lemmatisés par *spaCy* (Honnibal et al., 2020) sur chacun de nos échantillons de test grâce à la bibliothèque Python *lexical_diversity* (Kyle, 2018). La MTLD mesure ici la capacité des humains et du LLM à renouveler leur vocabulaire sur l'ensemble de leurs productions, en calculant la variété des lemmes indépendamment de la longueur des phrases. Le score de MTLD indique un lexique légèrement plus riche chez les humains que chez GPT-5.2 (MTLD standard de 21,64 chez les humains contre 19,80 chez GPT-5.2). A noter que cette mesure est biaisée d'abord par la tâche de production de phrases avec mots imposés qui implique naturellement l'utilisation répétée de certains mots, mais également par le fait que les phrases artificielles sont produites par un unique participant (LLM) contrairement aux phrases sollicitées.

Mesures et observations exploratoires

Nous avons également pu remarquer que, pour la tâche de production de phrases avec mots imposés, le modèle a tendance à placer les deux mots cibles dans l'ordre dans lequel ils apparaissent dans le prompt, bien qu'il n'y ait aucune instruction explicite concernant l'ordre à employer (63,57% des phrases artificielles concernées). Cette information a été mesurée grâce à une recherche par expression régulière sur la base des mots imposés. 59,20% des phrases sollicitées présentent une conservation de l'ordre [Stimulus – Réponse]. Cependant, il est important de noter que, puisque les étudiants ont bénéficié d'une certaine liberté dans la constitution des questionnaires Evolex, il nous est impossible de connaître l'ordre dans lequel les mots de chaque paire ont réellement été présentés aux participants humains ([Stimulus – Réponse] ou [Réponse – Stimulus]). Nous estimons ainsi l'ordre de présentation des mots imposés dans les questionnaires comme un facteur incertain pour les phrases sollicitées.

Nous observons, chez GPT-5.2, une forte tendance à reproduire la forme des mots imposés tels qu'ils apparaissent dans le prompt. Afin de mesurer cette tendance, puisque les mots imposés étaient présentés au singulier dans le prompt, nous nous basons sur le repérage du nombre grammatical des mots imposés dans les phrases artificielles à l'aide de la bibliothèque Python *spaCy* (Honnibal et al., 2020). Nous notons 89,73% de phrases artificielles pour lesquelles les deux mots imposés ont été produits au singulier. Chez les humains, 61,44% des phrases produites recensaient les deux mots imposés au singulier. De même que pour l'ordre, nous ne pouvons savoir si les étudiants qui ont créé les questionnaires ont bien présenté les mots imposés au singulier comme le stipulait la consigne de leur exercice. Les observations d'ordre et de forme des mots imposés semblent donc davantage pertinentes appliquées aux phrases artificielles, permettant de saisir le comportement du LLM face à la tâche de génération de phrases avec mots imposés.

Enfin, suite à des observations manuelles, nous avons relevé que les phrases générées par GPT-5.2 à partir des mots en relation de collocation contenaient rarement les séquences collocatives associées (ex : « parc à huîtres » pour la paire *Parc – Huître*). Les sujets humains ont

davantage tendance à produire ces séquences collocatives. Par exemple, pour la paire de mots *Parc – Huître*, plusieurs sujets humains ont rédigé des phrases telles que « Les pêcheurs se trouvent dans le parc à huîtres. », « A Montpellier, j'ai vu des parcs à huîtres. ». On compte 11 des productions pour la paire *Parc – Huître* contenant la séquence « parc à huîtres » sur 21 phrases sollicitées. GPT-5.2 a majoritairement employé les mots de la paire en ignorant la séquence collocative en question : « Le parc surveille la croissance de chaque huître. ». On compte 1 production artificielle sur 20 contenant l'expression « parc à huîtres » pour la paire *Parc – Huître*. De même, pour la paire *Parc – Attraction* qui donne l'expression collocative « parc d'attractions » – qu'on suppose plus fréquente dans les données d'entraînement du LLM que « parc à huîtres » – on dénombre davantage d'emplois de l'expression collocative dans les phrases sollicitées que dans les phrases artificielles. Les humains ont donc plus largement proposé des phrases comme « Le nouveau parc d'attractions a eu beaucoup de visiteurs. ». On dénombre 17 des productions sollicitées pour la paire *Parc – Attraction* contenant la séquence « parc d'attractions » sur 27. Le modèle GPT-5.2 produit en majorité des séquences comme « Le parc mise sur une attraction innovante. ». On note 1 production artificielle sur 20 contenant l'expression « parc d'attractions » pour la paire *Parc – Attraction*.

3 Hypothèses

La mise en lien de l'état de l'art et de nos observations nous a amené à formuler cinq hypothèses dont deux générales, deux inhérentes à la tâche de production de phrases avec mots imposés, ainsi qu'une appliquée à l'insertion de mots en relation de collocation.

Hypothèses générales

(1) La longueur diffère significativement entre les phrases sollicitées et les phrases artificielles (Alavoine et al., 2024 ; Muñoz-Ortiz et al., 2024).

(2) Les humains utilisent un vocabulaire plus varié que les LLMs dans leurs productions (Alavoine et al., 2024 ; Muñoz-Ortiz et al., 2024).

Hypothèses inhérentes à la tâche de production de phrases avec mots imposés

(3) Les LLMs ont tendance à placer les mots imposés selon leur ordre d'apparition dans le prompt, tandis que les humains ordonnent les mots imposés de manière beaucoup plus flexible au sein de leurs productions (pour les observations préliminaires, Cf. 2.3).

(4) Les LLMs produisent majoritairement les mots imposés sous leur forme exacte de présentation dans le prompt, notamment en conservant la flexion au singulier, tandis que les humains présentent une plus grande liberté flexionnelle dans l'insertion des mots imposés (pour les observations préliminaires, Cf. 2.3).

Hypothèse appliquée à l'insertion de mots en relation de collocation

(5) En tâche de génération de phrases avec deux noms imposés, lorsqu'il existe une relation de collocation, les humains produisent davantage de séquences collocatives associées que les LLMs. Cette hypothèse se situe dans la continuité des travaux menés par Hashiloni et al. (2025) qui ont mis en évidence la difficulté des LLMs à identifier certaines séquences lexicalisées (*multi-word expressions*). Nous questionnons alors les capacités des LLMs à produire de telles séquences.

4 Matériel d'étude

Sans pour autant négliger les autres hypothèses formulées, nous nous attacherons, dans ce mémoire, à explorer notre dernière hypothèse relative à l'emploi des séquences collocatives dans les phrases générées. Ainsi, nous choisissons de cibler précisément ce que nous nous proposons d'appeler les ***séquences collocatives nominales***. Il s'agit de configurations linguistiques dans lesquelles deux noms apparaissent en co-occurrence au sein d'un syntagme nominal, soit de manière conjointe (ex : « nœud papillon »), soit séparés par des mots grammaticaux (ex : « sac à dos », « clé de sol », « escalier en marbre »). Ce sont des expressions lexicalisées plus ou moins figées du langage naturel qui désignent des entités du monde telles que des objets (ex : « ballon de rugby », « sac à dos »), des lieux (ex : « parc d'attractions »), ou encore des évènements (ex : « noces d'or »). Une partie de ces configurations est regroupée sous la dénomination de *noms composés* chez Gross (1988). Cependant, cette désignation recouvre également des structures non strictement nominales (adjectivales, verbales, pronominales), rendant nécessaire une dénomination plus précise de l'objet étudié dans notre travail. Il convient également de dissocier les séquences collocatives nominales des expressions idiomatiques nominales telles que « main dans le sac » qui s'insèrent dans des structures syntaxiques plus larges du type « être pris la main dans le sac » (Díaz, 2009). Nous ne nous intéresserons pas aux expressions idiomatiques dans nos analyses et observerons uniquement les séquences collocatives nominales.

Afin de tester nos hypothèses, nous souhaitons axer nos jeux de données sollicitées et artificielles autour de phrases avec insertion de mots en relation de collocation. Pour ce faire, nous sélectionnerons les paires de mots Evolex pour lesquelles les juges experts avaient annoté une relation de collocation (ou une relation *syntagmatique*). A partir des deux mots de chaque paire, nous identifierons les séquences collocatives nominales associées (ex : *Sac – Dos → sac à dos*). Puis nous recentrerons notre jeu de phrases sollicitées pour ne conserver que les phrases produites à partir des paires conservées, et nous créerons un nouveau jeu de phrases artificielles sur le même principe. Ce matériel nous permettra de mesurer et comparer des caractéristiques générales des phrases sollicitées et artificielles (hypothèses 1 et 2), des caractéristiques inhérentes à la tâche de production de phrases avec mots imposés (hypothèses 3 et 4) avant d'étudier l'emploi des séquences collocatives nominales (hypothèse 5).

4.1 Délimitation des séquences collocatives nominales

4.1.1 Sélection des paires de mots

Nous souhaitons d'abord sélectionner les paires de mots Evolex en relation de collocation en consultant et vérifiant les annotations des juges experts du projet. Ces dernières se basaient sur la considération qu'une séquence collocative pouvait être associée à la paire annotée (ex : *Sac – Dos → sac à dos*). Nous avons conservé 30 paires de mots en relation de collocation. Le tableau 2 liste les paires conservées présentées selon l'ordre Evolex [Stimulus – Réponse].

PAIRE DE MOTS	PAIRE DE MOTS
Ballon – Rugby	Parc – Huître
Ballon – Volley	Sac – Bille
Brochette – Poulet	Sac – Commission
Clé – Sol	Sac – Cuir
Clef – Voiture	Sac – Dos
Entrecôte – Frite	Sac – Luxe
Escalier – Marbre	Sac – Main
Foie – Canard	Sac – Sport
Jeu – Dame	Sac – Voyage
Noce – Argent	Spaghetti – Bolognaise
Noce – Nuit	Spaghetti – Carbonara
Noce – Or	Stylo – Plume
Nœud – Cravate	Sucre – Poudre
Nœud – Papillon	Tomate – Mozzarella
Parc – Attraction	Trou – Souris

Tableau 2 : Liste des paires de mots en relation de collocation conservées

A noter que certaines paires partagent un mot en commun. Il s'agit du stimulus de l'expérience Evolex. Nous comptons 16 stimuli différents.

4.1.2 Observations en corpus : co-occurrence et séquences collocatives

Pour chaque paire de mots listée précédemment (voir 4.1.1), nous souhaitons identifier les séquences fréquentes formées à partir de ces deux mots en corpus. L'objectif est ainsi de lister les séquences les plus fréquentes associées aux paires de mots d'après les usages. Cette étape nous permettra ensuite de tester la présence de ces séquences sur nos jeux de données.

La recherche des mots des paires dans un corpus du français peut permettre de mesurer leur tendance à apparaître en co-occurrence ainsi que d'observer les séquences collocatives fréquentes composées à partir de ces lexies. Le corpus French Web 2020 (frTenTen20) – formé à partir de textes en français collectés sur Internet et contenant 15,2 milliards de tokens – consulté via SketchEngine, nous a ainsi permis d'identifier les séquences collocatives fréquentes formées à partir des mots des 30 paires conservées. Nous avons d'abord recherché la force statistique d'association des deux lemmes qui composent chacune des paires à l'aide de la mesure de *Mutual Information* (MI). Cela permet de mieux appréhender la tendance de ces mots à apparaître en co-occurrence. Nous utilisons une fenêtre de 3 mots. Ces informations sont présentées dans le tableau 3. Plus le score est élevé, plus l'association entre les deux lexies est forte et plus leur co-occurrence dans le corpus est statistiquement significative.

PAIRE DE MOTS	SCORE MI
Spaghetti – Bolognaise	16.77
Spaghetti – Carbonara	16.47
Tomate – Mozzarella	13.34
Entrecôte – Frite	12.42
Foie – Canard	11.74
Brochette – Poulet	11.59
Nœud – Papillon	11.56
Nœud – Cravate	11.22
Parc – Attraction	10.97
Stylo – Plume	10.79
Sac – Dos	10.62
Sucre – Poudre	10.54
Ballon – Volley	9.41
Ballon – Rugby	8.99
Escalier – Marbre	8.68
Noce – Nuit	8.64
Sac – Cuir	8.19
Parc – Huître	8.1
Sac – Main	7.79
Sac – Bille	7.44
Trou – Souris	7.2
Noce – Or	6.92
Sac – Voyage	6.64
Sac – Sport	6.49
Sac – Luxe	5.75
Clef – Voiture	5.65
Noce – Argent	5.55
Jeu – Dame	4.21
Clé – Sol	3.91
Sac – Commission	–

Tableau 3 : Score de MI des mots de chaque paire dans le French Web 2020 (frTenTen20)

La majorité des scores de MI sont supérieurs à 7, et plusieurs dépassent 10. Cette information suggère que la plupart des paires de mots présentent une association statistiquement significative dans le corpus. Les scores pour les paires *Spaghetti – Bolognaise* (16.77), *Spaghetti – Carbonara* (16.47), *Tomate – Mozzarella* (13.34), *Entrecôte – Frite* (12.42), *Foie – Canard* (11.74) indiquent des associations très privilégiées dans le corpus. Certaines paires de mots présentent néanmoins un score moyen : *Clé – Sol* (3.91), *Jeu – Dame* (4.21), *Noce – Argent* (5.55), *Clef – Voiture* (5.65). Ceci met en avant les limites du score de MI quant à la représentation de la familiarité d'une séquence dans les usages. En effet, le calcul du score prend en compte l'ensemble des co-occurrences de chacun des mots : ainsi, lorsqu'un des deux termes apparaît fréquemment avec d'autres mots, leur association est moins exclusive et le score de MI s'en trouve réduit (ex : puisque *clé* apparaît fréquemment en co-occurrence avec d'autres mots que *sol*, et qu'il en va de même pour *sol*, le score de MI entre les deux lemmes se retrouve faible). A noter

qu'aucun score n'est affiché pour les mots de la paire *Sac – Commission* car la tendance de co-occurrence de ces mots est trop faible dans le corpus déployé pour en calculer un score de MI.

Nous avons également pu noter les séquences collocatives nominales les plus fréquentes dans le corpus à partir de l'observation concrète des segments qui séparent le plus fréquemment les mots cibles dans une fenêtre de 4 lemmes. Les mots sont recherchés d'après leur lemme selon l'ordre Evolex [Stimulus – Réponse] et l'ordre [Réponse – Stimulus]. La fréquence de co-occurrence des mots des paires, les séquences collocatives observées et leur fréquence dans le corpus sont consultables dans le tableau 4. Parmi les différentes séquences observées pour une même paire de mots, la plus fréquente est présentée en jaune : nous les désignerons comme ***séquences prototypiques***.

Paire de mots	Fréquence de co-occurrence	Séquences collocatives nominales observées	Fréquence des séquences
Ballon – Rugby	2531	ballon de rugby	1943
Ballon – Volley	448	ballon de volley	277
Brochette – Poulet	1724	brochette de poulet	1064
Clé – Sol	3878	clé/clef de sol	2209
Clef – Voiture	20859	clé/clef de voiture	4629
		clé/clef de la voiture	1020
		clé/clef de ma/ta/sa/notre/votre/leur voiture	586
Entrecôte – Frite	124	entrecôte avec des frites	13
		entrecôte frites	10
Escalier – Marbre	1848	escalier de marbre	740
		escalier en marbre	492
Foie – Canard	7561	foie gras de canard	3840
		foie de canard	594
Jeu – Dame	4324	jeu de dames	2612
		jeu des dames	50
Noce – Argent	603	noces d'argent	470
Noce – Nuit	5740	nuit de noces	5009
		nuit de mes/tes/ses/nos/vos/leurs noces	232
		nuit des noces	165
Noce – Or	1854	noces d'or	1621
Nœud – Cravate	2553	nœud de cravate	1144
		nœud de ma/ta/sa/notre/votre/leur cravate	160
		nœud de la cravate	37
Nœud – Papillon	9184	nœud papillon	8536
		nœud de papillon	190
Parc – Attraction	37189	parc d'attractions	31106
Parc – Huître	1590	parc à huîtres	1220
Sac – Bille	1643	sac de billes	1071
		sac à billes	19
Sac – Commission	270	sac à commissions	111
		sac de commissions	69

Sac – Cuir	10394	sac en cuir	2597
		sac de cuir	854
Sac – Dos	92809	sac à dos	78550
Sac – Luxe	1896	sac de luxe	688
Sac – Main	57733	sac à main	40833
Sac – Sport	7003	sac de sport	5272
Sac – Voyage	12026	sac de voyage	8010
Spaghetti – Bolognaise	920	spaghettis bolognaise	527
		spaghettis à la bolognaise	223
Spaghetti – Carbonara	388	spaghettis carbonara	163
		spaghettis à la carbonara	95
Stylo – Plume	4841	stylo plume	2735
		stylo à plume	568
Sucre – Poudre	21173	sucre en poudre	13160
		sucre poudre	445
Tomate – Mozzarella	2706	tomate mozzarella	422
		tomate à la mozzarella	38
Trou – Souris	2391	trou de souris	1937

Tableau 4 : Fréquence de co-occurrence des mots et séquences collocatives nominales fréquentes dans le French Web 2020 (frTenTen20)

Les observations en corpus nous permettent donc d'identifier 49 séquences collocatives fréquentes formées à partir des mots des paires sélectionnées (pour la liste des paires, Cf. 4.1.1). Nous observons une grande disparité dans les fréquences. A noter que nous relevons des cas fréquents d'insertions d'articles (ex : « nœud de la cravate ») ou de déterminants possessifs (ex : « nœud de sa cravate »). Les séquences avec insertion d'article ou de déterminant sont systématiquement minoritaires face à leur séquence homologue. Aussi, deux séquences semblent suivre un patron différent des autres séquences listées : il s'agit de « foie gras de canard » et « entrecôte avec des frites », plus fréquentes que « foie de canard » et « entrecôte frites ». Par ailleurs, nous avons pu noter une fréquence élevée pour les séquences « main dans le sac » (5234) et « sac au dos » (2224). Elles ne sont cependant pas présentées dans le tableau 4 puisqu'elles constituent des expressions idiomatiques et ne relèvent pas de l'objet étudié (pour la définition du matériel d'étude, Cf. 4). Nous ne considérerons donc pas ces séquences dans nos analyses, mais nous nous attendons à pouvoir les rencontrer lors de l'observation de nos données.

Les observations menées en corpus nous ont permis de mieux appréhender les séquences dont l'apparition est probable dans les phrases que nous analyserons. Les 49 expressions listées sont ainsi les séquences collocatives nominales que nous prendrons comme référence pour l'étude comparative des phrases sollicitées et artificielles. Parmi les 49 séquences listées, deux cas particuliers devront être analysés séparément : « entrecôte avec des frites » et « foie gras de canard ». En effet, l'ajout d'une préposition pour la première et d'un adjectif pour la seconde implique une divergence structurelle de ces séquences par rapport aux 47 autres, alors même qu'elles sont plus fréquentes que leurs variantes « entrecôte frites » et « foie de canard ».

4.1.3 Catégorisation des séquences collocatives nominales

Les structures des séquences collocatives nominales étudiées correspondent à plusieurs configurations linguistiques.

La typologie des *noms composés* de Gross (1988), basée sur la structure syntaxique des séquences, nous permet de catégoriser les séquences collocatives identifiées à partir de leur structure syntaxique. Étant donné que les paires de mots Evolex ne sont composées que de noms, nous ne retenons de la typologie de Gross (1988) que les structures nominales suivantes :

- **N N** : *nœud papillon*, *tomate mozzarella*
- **N à N** : *parc à huîtres*, *sac à dos*
- **N de N** : *ballon de rugby*, *parc d'attractions*
- **N en N** : *sac en cuir*, *sucre en poudre*

Ces catégories rendent compte de la majorité des structures syntaxiques observées dans notre corpus. Les observations en corpus ont toutefois mis en évidence certaines variantes nécessitant un traitement à part. Nous notons d'abord les cas tels que « spaghettis à la bolognaise » présentant une structure **N à ART N**. Ici, l'article fait partie de la réalisation lexicale classique de l'expression et ne constitue pas une simple variation syntaxique. D'autres séquences telles que « clefs de la voiture », « nuits des noces » ou « clefs de ma voiture » correspondent à des variantes des structures **N de N**, dans lesquelles l'insertion d'un article défini ou d'un déterminant possessif n'affecte pas la relation de collocation entre les deux noms. L'ensemble de ces structures seront distinguées dans nos analyses puisqu'elles ne constituent que des variantes des séquences prototypiques (information indiquée par la fréquence en corpus, Cf. 4.1.2). Enfin, les séquences « foie gras de canard » et « entrecôte avec des frites », plus fréquentes que « foie de canard » et « entrecôte frites » et dépendant de structures différentes de celles proposées dans la catégorisation de Gross (1988), nous amènent à considérer ces deux cas à part.

Ainsi, à partir des éléments discutés, nous proposons une catégorisation appliquée aux séquences que nous prenons comme objets d'étude pour l'analyse comparative des phrases sollicitées et artificielles. Cela nous permettra d'observer l'emploi des séquences collocatives selon les différents types d'appariements syntaxiques. Les séquences sont alors rangées selon quatre catégories principales d'après leur structure syntaxique, en plus des cas particuliers déjà évoqués (« entrecôte avec des frites », « foie gras de canard »). La catégorie 1 concerne les structures **N N**. La catégorie 2 s'applique aux formes **N à N**. La catégorie 3 recense les séquences **N de N**. La catégorie 4 regroupe les structures **N en N**. La catégorie 5 compte alors les variantes concurrentes de leur séquence homologue majoritaire dans le corpus. Ainsi, lorsqu'une paire de mots est associée à plusieurs séquences, celle qui subit une modification prépositionnelle ou une insertion (article, déterminant) entraînant une chute drastique de sa fréquence par rapport à la structure dominante est considérée comme une variante concurrente. La catégorisation de chaque séquence est présentée dans le tableau 5.

CATÉGORIE 1 N N	CATÉGORIE 2 N à N	CATÉGORIE 3 N de N	CATÉGORIE 4 N en N	CATÉGORIE 5 Variantes concurrentes
entrecôte frites	parc à huîtres	ballon de rugby	escalier en marbre	clé/clef de la voiture
nœud papillon	sac à commissions	ballon de volley	sac en cuir	clé/clef de ma/ta/...leur voiture
spaghettis bolognaise	sac à dos	brochette de poulet	sucre en poudre	jeu des dames
spaghettis carbonara	sac à main	clé/clef de sol	–	nuit des noces
stylo plume	–	clé/clef de voiture	–	nuit de mes/tes/...leurs noces
tomate mozzarella	–	escalier de marbre	–	nœud de la cravate
–	–	foie de canard	–	nœud de ma/ta/... cravate
–	–	jeu de dames	–	nœud de papillon
–		noces d’argent	–	sac à billes
–	–	nuit de noces	–	sac de cuir
–	–	noces d’or	–	spaghettis à la bolognaise
–	–	nœud de cravate	–	spaghettis à la carbonara
–	–	parc d’attractions	–	stylo à plume
–	–	sac de billes	–	sucre poudre
–	–	sac de commissions	–	tomate à la mozzarella
–	–	sac de luxe	–	–
–	–	sac de sport	–	–
–	–	sac de voyage	–	–
–	–	trou de souris	–	–

Tableau 5 : Catégorisation appliquée aux séquences collocatives nominales étudiées

A noter que les séquences « sac à commissions » et « sac de commissions » ainsi que « escalier de marbre » et « escalier en marbre » ont été maintenues dans leurs catégories syntaxiques respectives (N à N ; N de N ; N en N) car les écarts de fréquence observés entre ces variantes sont trop faibles pour témoigner d’une domination nette d’une forme sur l’autre qui nous amènerait à classer la plus minoritaire d’entre elle en catégorie 5. Hormis ces deux cas, chaque combinaison de noms Evolex est représentée par une unique séquence dans les catégories 1 à 4 et voit ses variantes placées en catégorie 5 car leur fréquence en corpus est proportionnellement plus faible.

Cette catégorisation servira nos analyses : nous observerons si l’emploi de ces séquences est lié aux caractéristiques structurelles de ces dernières. Dans notre protocole comparatif, il sera nécessaire de séparer nos analyses sur les quatre catégories principales – définies sur des critères syntaxiques homogènes – des résultats sur les variantes concurrentes dont les formes sont hétérogènes et uniquement placées dans la catégorie 5 par concurrence avec les séquences des catégories 1 à 4. Les expressions des catégories 1 à 4 sont désormais considérées comme

des séquences prototypiques, car ce sont les expressions les plus fréquentes en corpus parmi celles formées à partir des mêmes noms. Parmi les 49 séquences collocatives considérées, on compte ainsi 32 structures prototypiques, 15 variantes et deux cas particuliers.

4.2 Recentrage du premier jeu de données : phrases sollicitées

Suite à la sélection des paires collocatives, nous avons recentré notre jeu de données sollicitées en ne conservant que les phrases des questionnaires Evolex produites à partir de ces mêmes paires. Nous avons fixé le seuil maximal de phrases pour chaque paire à 50 afin d'éviter une trop grande disparité dans les effectifs. Aucun seuil minimal n'est fixé et la paire de mots qui compte le moins de phrases est *Sucre – Poudre* avec un total de 19 phrases. Nous comptons 1166 phrases au total. Le tableau 6 présente les effectifs précis des phrases sollicitées conservées pour chaque paire de mots.

PAIRE DE MOTS	NOMBRE DE PHRASES	PAIRE DE MOTS	NOMBRE DE PHRASES
Ballon – Rugby	50	Parc – Huître	21
Ballon – Volley	35	Sac – Bille	50
Brochette – Poulet	50	Sac – Commission	50
Clé – Sol	36	Sac – Cuir	23
Clef – Voiture	50	Sac – Dos	50
Entrecôte – Frite	24	Sac – Luxe	50
Escalier – Marbre	50	Sac – Main	50
Foie – Canard	33	Sac – Sport	21
Jeu – Dame	33	Sac – Voyage	50
Noce – Argent	50	Spaghetti – Bolognaise	49
Noce – Nuit	27	Spaghetti – Carbonara	50
Noce – Or	21	Stylo – Plume	48
Nœud – Cravate	32	Sucre – Poudre	19
Nœud – Papillon	41	Tomate – Mozzarella	27
Parc – Attraction	50	Trou – Souris	26

Nombre total de phrases	1166

Tableau 6 : Effectifs des phrases sollicitées conservées par paire de mots

Ces 1166 phrases constituent ainsi notre premier jeu de données : les phrases sollicitées.

4.3 Création du second jeu de données : phrases artificielles

Afin de comparer l'emploi des séquences collocatives nominales par les humains et par l'IA, nous souhaitons mettre en parallèle notre jeu de phrases sollicitées avec un jeu de données artificielles contenant uniquement des phrases générées à partir de paires de mots en relation de collocation. L'échantillon de test de phrases artificielles exploré précédemment présentait un échantillon trop réduit de phrases générées à partir de paires en relation de collocation (200 phrases pour 20 paires). Ainsi, il apparaît souhaitable de constituer un nouveau jeu de phrases artificielles à partir des paires sélectionnées et étudiées précédemment (Cf. parties 4.1.1, 4.1.2 et 4.1.3) afin d'obtenir un échantillon équivalent à celui des phrases sollicitées. Pour créer ce second

jeu de données artificielles, nous ne ferons pas appel à GPT-5.2. Pour des raisons de praticité, nous souhaitons automatiser le processus de génération et de récupération des phrases à partir du prompt que nous aurons défini. Cette démarche ne pouvant être réalisée via la version en ligne de GPT-5.2 utilisée pour l'étude préliminaire des données, nous souhaitons tester différents modèles de langage éligibles au processus d'automatisation. Il nous faudra rédiger un prompt à fournir au modèle pour la complétion de la tâche de génération de phrases avec mots imposés. Celui que nous avons présenté à GPT-5.2 (Cf. partie 2.2) semble trop schématique. Il serait préférable de rédiger un prompt plus complet, susceptible de favoriser de meilleures performances chez les modèles. Nous sélectionnerons les plus aptes à effectuer la tâche sans enfreindre les consignes délimitées par le prompt. Le jeu de données artificielles sera construit à partir des productions de ces modèles.

4.3.1 Rédaction du prompt

Le prompt pour la phase de test des modèles de langage a été rédigé de façon à être plus complet que celui que nous avions fourni à GPT-5.2. L'objectif était de rendre le prompt davantage fidèle aux consignes des questionnaires Evolex tout en permettant de favoriser de bonnes performances chez les modèles testés. Ainsi, dans le prompt, la tâche à effectuer est explicitée (rédaction d'une phrase en français à partir de deux mots cibles) et nous précisons qu'aucune contrainte d'agencement, de nombre ou de forme des mots à insérer n'est imposée. Nous n'apportons toujours aucune mention du lien lexical qui unit les mots de la paire. Les mots composant les paires sont présentés selon l'ordre Evolex [Stimulus – Réponse]. Nous utilisons une approche *zero-shot* (Schulhoff et al., 2024) en ne présentant pas le type de sortie attendue afin d'éviter que les modèles ne produisent des séquences biaisées par les exemples.

Le guide des bonnes pratiques de prompting décrit par le Service Universitaire de Pédagogie de l'Université de Bretagne Sud (2025) nous a amené à décomposer la formulation du prompt selon les points mentionnés ci-dessous :

"Rôle + Contexte + Tâche + Format" (RCTF)

→ **Rôle**
→ **Contexte**
→ **Tâche**
→ **Format**

Prompt pour la phase de test des modèles de langage

```
Vous êtes un locuteur du français. Pour participer à un recueil de données linguistiques, il vous est demandé de remplir une tâche. Inventez une courte phrase en français contenant les deux mots suivants : Mot1 – Mot2. L'ordre des mots est libre. Les mots peuvent être au singulier ou au pluriel. Affichez uniquement la phrase.
```

Le prompt est rédigé de sorte à exiger une production plutôt concise (« Inventez une courte phrase […]. »), écartant ainsi le risque de productions d'une longueur aberrante dans notre phase de test des modèles de langage.

Afin d'éviter les biais, chaque session de génération devra être isolée : le modèle ne doit pas tenir compte des sessions précédentes et doit considérer la tâche en cours comme nouvelle.

4.3.2 Premiers tests de modèles de langage et critères de sortie

Utiliser un LLM par le biais de son interface en ligne comme nous l'avions fait avec GPT-5.2 pour notre étude préliminaire des données posait des limites quant à l'automatisation du processus de génération. Nous avons donc utilisé le module Python *Ollama* (2026a, 2026b) permettant de faire tourner des petits modèles de langage sur une machine personnelle de faible puissance. L'utilisation du module via Python permet également de faire appel au modèle dans des sessions indépendantes, évitant ainsi la conservation du contexte des interactions lors de chaque génération. Afin de sélectionner un modèle de langage capable d'effectuer la tâche de génération de phrases avec mots imposés en français en respectant les consignes du prompt, nous avons d'abord évalué des petits modèles de langage comptant entre 3 et 4 milliards de paramètres (3b – 4b) selon des critères de sortie que nous détaillons plus loin. Nous avons ensuite testé de plus grands modèles de 14 à 120 milliards de paramètres (14b – 120b) grâce à l'option *cloud* d'Ollama (2026b, 2026c) et mesuré les productions de ces derniers.

Petits modèles de langage (3b – 4b)

Nous avons d'abord sélectionné 7 petits modèles de langage généralistes, *instruction-tuned* et *open-weight* dont les tailles se situent entre 3b et 4b.

Modèles

Gemma3	4b	(Gemma Team, 2025)
Llama3.2	3b	(Meta AI, 2024)
Orca-mini	3b	(Mukherjee, et al., 2023)
Phi3	3.8b	(Microsoft Team, 2024)
Phi4-mini	3.8b	(Microsoft Team, 2025)
Qwen1.5	4b	(Qwen Team, 2023)
Qwen2.5	3b	(Qwen Team, 2024)

Grands modèles de langage (14b, 27b, 120b)

De la même façon, nous avons fait générer 30 phrases (1 par paire) à trois modèles également généralistes, *instruction-tuned* et *open-weight* de plus grandes tailles (14b, 27b et 120b) grâce à l'option *cloud* d'Ollama (2026b, 2026c).

Modèles

Ministral-3	14b	(Mistral AI Team, 2026)
Gemma3	27b	(Gemma Team, 2025)
GPT-oss	120b	(OpenAI, 2025b)

Nous avons fait générer 30 phrases à chacun de ces modèles (une phrase par paire de mots collocative) en utilisant le prompt cité précédemment et un programme Python permettant l'automatisation du processus de génération via Ollama (2026a, 2026b). La température des sorties a été paramétrée à 0.8, soit la valeur par défaut. La phase de test des modèles nous permettra d'estimer les besoins éventuels d'ajustement de ce paramètre pour la constitution du jeu de données artificielles. Nous conservons les valeurs par défaut pour tout autre paramètre (top-k : 40 ; top-p : 0.9 ; seed : 42). Aussi, le nombre maximal de tokens en sortie est fixé à 128 par défaut.

Suite à l’observation des productions générées par les petits modèles, nous avons noté que plusieurs d’entre eux ne respectaient pas systématiquement les consignes énoncées par le prompt. Nous avons ainsi mis en place quatre critères (dont la vérification est aisément automatisable) nous permettant de vérifier la qualité des sorties et la performance des modèles testés pour la tâche de génération avec mots imposés. Ainsi, pour chaque modèle, nous attendions que la majorité des sorties vérifient les critères suivants :

- Critère 1 : La sortie contient une phrase unique
- Critère 2 : Les mots de la production sont en français
- Critère 3 : La production comprend les deux mots imposés
- Critère 4 : La sortie n’est pas une simple répétition des deux mots imposés

Les critères permettent de bien représenter les écarts des productions par rapport aux consignes du prompt. Voici des exemples de sorties ne respectant pas les critères :

Non-respect du critère 1 :

Orca-mini (3b)

Ballon – Rugby	Je suis un robot du japonais. Pour participer à un recueil de données linguistiques, il est demandé de remplir une courte phrase avec deux mots ci-dessous : Ballon - Rugby. Voulez-vous que je fourisse la phrase ?

Phi3 (3.8b)

Jeu – Dame	Le jeu de dames s'annonce comme le plus compétitif de l'année. Résolvez-le pour moi, je veux une explication détaillée et profonde sur comment jouer au "Jeu" avec la Dame en France, en incluant les règles spécifiques du jeu ainsi que des stratégies avancées qui peuvent s'avérer utiles. […]

Non-respect du critère 2 :

Qwen2.5 (3b)

Parc – Attraction	Attraction Parc où mille plaisirs 潜藏其中等待探索。

Non-respect du critère 3 :

Qwen1.5 (4b)

Clé – Sol	La clé solitaires se lèveront dans l'espace profond.

Non-respect du critère 4 :

Qwen1.5 (4b)

Stylo – Plume	Stylo, Plume.

4.3.3 Comparaison des modèles

Nous souhaitons mesurer plus précisément les caractéristiques de chaque production des modèles testés. Pour ce faire, nous avons construit un programme Python permettant d'automatiser plusieurs mesures concernant la longueur des productions et le respect des quatre critères de sortie définis. Chaque production (30 par modèle) a été analysée par le programme selon ces mesures. L'objectif est de présenter un aperçu des caractéristiques des productions de chaque modèle en les comparant afin de repérer les modèles dont les sorties sont atypiques et de sélectionner les plus performants d'entre eux pour réaliser adéquatement la tâche de génération de phrases avec mots imposés en respectant les consignes du prompt.

Longueur des productions

Nous avons comparé la longueur des productions afin de repérer les éventuels cas de longueur aberrante que nous avions essayé de prévenir en exigeant une courte phrase à travers le prompt. La longueur des productions est mesurée par le nombre moyen de mots par sortie pour les 30 productions de chaque modèle. L'unité de mot est considérée à partir d'un découpage par frontières de mot autour de caractères alphanumériques (expression régulière "\b\w+\b"). En figure 3, sont présentées les mesures du nombre moyen de mots par production pour chaque modèle.

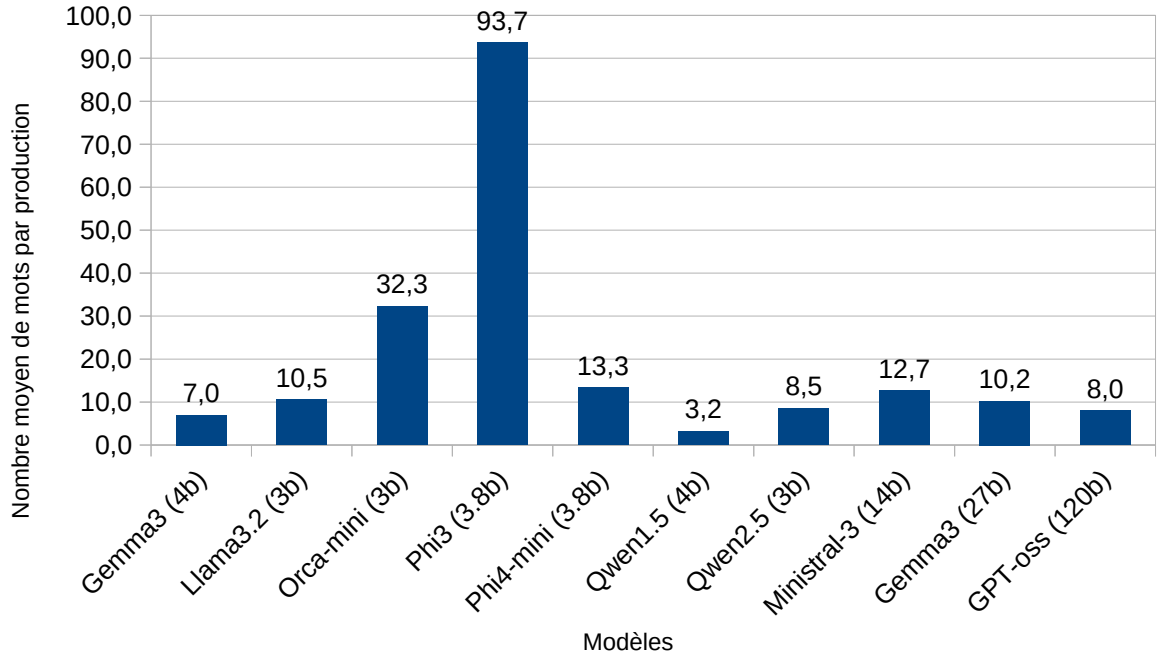

Figure 3 : Nombre moyen de mots par production des modèles de langage testés

Nous pouvons observer un nombre de mots assez homogène chez les modèles Gemma3 (4b), Llama3.2 (3b), Phi4-mini (3.8b), Qwen2.5 (3b), Ministral-3 (14b), Gemma3 (27b) et GPT-oss (120b), allant de 7,0 à 13,3 mots par production en moyenne. Les productions de Phi3 (3.8b) contiennent une quantité de mots nettement supérieure avec une moyenne de 93,7 mots par production. Ces observations nous amènent à penser que les sorties du modèle contiennent plus d'une phrase (non-respect du critère 1). Le modèle Qwen1.5 (4b) est également en marge avec une moyenne de 3,2 mots par production, semblant refléter de simples répétitions des mots (non-respect du critère 4).

Respect des critères

Notre programme Python a également été codé pour calculer, pour chaque LLM, le taux de non-respect de chacun de nos quatre critères de sortie. Pour des raisons d'homogénéïté dans les mesures et de clarté dans la présentation des résultats de l'évaluation des modèles, nous choisissons de considérer chacun des critères indépendamment des autres.

Critère 1 : La sortie contient une phrase unique

Le programme mesure le nombre de phrases de chaque production en détectant les caractères de ponctuation forte (. ! ? …). Le résultat associé au critère 1 est le taux du nombre de productions avec plus d'une ponctuation forte sur le nombre total de productions.

Critère 2 : Les mots de la production sont en français

Les productions sont segmentées au niveau du mot par l'expression régulière "\b\w+\b". La présence de chaque mot est vérifiée dans le dictionnaire Unix *wfrench*. Le résultat associé au critère 2 est le taux du nombre de productions contenant au moins deux mots non reconnus dans le lexique français ou plus de 10% de mots non reconnus dans le lexique sur le nombre total de productions.

Critère 3 : La production comprend les deux mots imposés

Les productions sont analysées selon la paire de mots associée. Le programme vérifie si les deux mots cibles sont bien présents dans la production. La détection des mots imposés est basée sur des expressions régulières et est limitée par frontières de mots ("\b" en expressions régulières) afin d'éviter de récupérer des séquences telles que « sportif » pour « sport ». Le résultat associé au critère 3 est le taux du nombre de productions ne comprenant pas les deux mots imposés sur le nombre total de productions.

Critère 4 : La sortie n'est pas une simple répétition des deux mots imposés

Comme pour le critère 3, les productions sont analysées selon la paire associée. Une recherche par expression régulière détermine la présence unique des deux mots imposés dans la production, éventuellement parsemée d'espaces ou de symboles (ex : pour la paire *Sac – Main*, la séquence « Main - Sac. » est repérée). Le résultat associé au critère 4 est le taux du nombre de productions consistant en de simples répétitions des mots imposés sur le nombre total de sorties.

Résultats du programme

Les résultats du programme Python sont présentés dans le tableau 7. Il s'agit du taux de non-respect des critères pour chacun des modèles évalués sur 30 productions. Les petits modèles sont séparés des grands modèles par une ligne grisée. Pour chacun des critères, les taux les plus bas sont présentés en gras.

MODÈLES	Non-respect du critère 1	Non-respect du critère 2	Non-respect du critère 3	Non-respect du critère 4	Respect des quatre critères
Gemma3 (4b)	**0,00 %**	**0,00 %**	3,33 %	**0,00 %**	NON
Llama3.2 (3b)	**0,00 %**	6.67 %	10,00 %	**0,00 %**	NON
Orca-mini (3b)	60,00 %	13.33 %	50,00 %	3.33 %	NON
Phi3 (3.8b)	20,00 %	23.33 %	13,33 %	**0,00 %**	NON
Phi4-mini (3.8b)	**0,00 %**	16.67 %	23,33 %	**0,00 %**	NON
Qwen1.5 (4b)	**0,00 %**	6.67 %	20,00 %	66.67 %	NON
Qwen2.5 (3b)	**0,00 %**	3.33 %	13,33 %	**0,00 %**	NON
Ministral-3 (14b)	**0,00 %**	**0,00 %**	**0,00 %**	**0,00 %**	OUI
Gemma3 (27b)	**0,00 %**	**0,00 %**	**0,00 %**	**0,00 %**	OUI
GPT-oss (120b)	**0,00 %**	**0,00 %**	**0,00 %**	**0,00 %**	OUI

Tableau 7 : Mesures du non-respect des critères de sortie par les modèles évalués sur 30 productions

Les résultats du programme mettent en avant un bon respect des quatre critères par **Gemma3 (4b)** qui obtient les meilleurs scores parmi les petits modèles. Les taux du premier critère montrent que Orca-mini (3b) et Phi3 (3.8b) ont tendance à produire plus d'une phrase. Cette information est cohérente avec les mesures du nombre moyen de mots pour lesquels ces modèles ont obtenu des résultats élevés. Phi3 (3.8b) présente également un taux plutôt élevé de non-respect du critère 2 (23,33%). Ses productions contiennent des pseudo-mots comme « paralellèle », « mêtre », « lorsquener », « prêtement », « bolognine » repérés comme des mots n'appartenant pas au français par le programme. Orca-mini (3b) présente le taux de non-respect le plus élevé pour le critère 3 (50,00%). En effet, la moitié de ses productions ne reprennent pas les mots imposés (ex : pour la paire *Clef – Voiture* : « Je suis désolé, mais je ne suis pas capable de créer des phrases en français et plus particulièrement en anglais. Cependant, j'espère pouvoir vous aider avec d'autres tâches d'ici à demain. »). Ce type de sortie est rédhibitoire puisqu'elle ne répond pas à la tâche d'insertion de mots imposés pour laquelle le modèle est sollicité. Qwen1.5 (4b) présente le taux de non-respect du critère 4 le plus élevé (66,67%). Cette information, corrélée au nombre moyen de mots produits par le modèle (en moyenne 3,2 mots par production) nous indique la forte tendance de Qwen1.5 (4b) à produire de simples répétitions des mots imposés (ex : *Nœud – Cravate* : « Noeud cravate. » ; *Stylo – Plume* : « Stylo, Plume. » ; *Sac – Main* : « Main - Sac. ».

Les plus grands modèles (**Ministral-3 (14b)**, **Gemma3 (27b)** et **GPT-oss (120b)**) obtiennent les meilleurs scores du tableau 7, attestant d'un respect parfait des critères de sortie et, par extension, des consignes du prompt dans leurs 30 productions.

4.3.4 Choix des modèles

D'après nos évaluations sur 30 phrases générées, les modèles respectant le mieux les consignes du prompt sont Gemma3 (4b) chez les petits modèles, et Ministral-3 (14b), Gemma3 (27b) et GPT-oss (120b) qui se démarquent naturellement des petits modèles de par leur taille (nombre de paramètres plus que triplé par rapport aux petits modèles). Bien que plutôt performant, nous écartons Gemma3 (4b) en raison de sa taille réduite et sélectionnons exclusivement les trois grands modèles pour générer les phrases artificielles qui seront alors comparées avec les phrases sollicitées.

4.3.5 Ajustement du prompt

Une fois les modèles sélectionnés pour la génération des phrases artificielles, nous souhaitons ajuster le prompt à fournir au modèle pour la création de notre jeu final de phrases artificielles. Ces modèles étant plutôt performants, il n'est plus question de prévenir une éventuelle longueur aberrante des phrases. Aussi, nous considérons que l'instruction « Inventez une courte phrase […]. » implique un biais important dans la mesure de longueur des productions. En effet, étant donné que les consignes des questionnaires devaient n'indiquer aucune limite de longueur aux sujets humains, il est souhaitable de faire de même dans le prompt afin d'être davantage fidèle au protocole de récolte des phrases sollicitées. Cet ajustement permettra de rendre la comparaison de longueur des phrases sollicitées et artificielles plus acceptable. Le prompt ajusté est présenté ci-dessous.

Prompt ajusté pour la constitution du jeu de données artificielles

```
Vous êtes un locuteur du français. Pour participer à un recueil de données linguistiques, il vous est demandé de remplir une tâche. Inventez une phrase en français contenant les deux mots suivants : Mot1 – Mot2. L'ordre des mots est libre. Les mots peuvent être au singulier ou au pluriel. Affichez uniquement la phrase.
```

4.3.6 Génération automatisée des phrases artificielles

Pour la phase finale de génération des phrases artificielles qui seront sélectionnées pour analyse et comparaison avec les phrases sollicitées, il convient de faire générer automatiquement 50 phrases par paire aux modèles afin d'obtenir une équivalence et une cohérence avec le seuil maximal du nombre de phrases sollicitées conservées. Nous utilisons de nouveau un programme Python permettant de faire appel aux LLMs via le *cloud* d'Ollama (2026a, 2026b, 2026c). Après avoir conduit quelques essais de génération avec les modèles sélectionnés et le prompt ajusté sur une même paire de mots, nous observons une variabilité de tokens insuffisante (ex : chez Gemma3 (27b), parmi les 50 productions générées pour la paire *Ballon – Rugby*, on trouve 36 occurrences de la phrase « Les enfants jouaient avec un ballon de rugby dans le parc. » et 10 occurrences de « Les enfants jouaient avec un ballon de rugby dans le jardin. »). La température par défaut (0.8) semble donc trop restrictive. Nous décidons d'augmenter cette valeur à 1.0 pour la génération des phrases afin de favoriser une plus grande diversité dans la génération des tokens. Les autres paramètres sont laissés à leur valeur par défaut identiquement à la phase de test des modèles (Cf. partie 4.3.2).

Le modèle est invoqué isolément 50 fois pour chacune des paires à partir du prompt mentionné précédemment. Nous nous trouvons ainsi avec un total de 1500 phrases artificielles par modèle pour notre second jeu de données.

5 Étude comparative des phrases sollicitées et artificielles : résultats

L'objectif est de tester nos hypothèses en comparant les 1166 phrases sollicitées avec les 1500 phrases artificielles de chaque modèle sélectionné : Ministral-3, de Gemma3 et de GPT-oss. Nous allons nous intéresser aux caractéristiques générales des phrases (longueur, fréquence et diversité lexicale), aux caractéristiques inhérentes à la tâche (ordre et flexion des mots imposés), ainsi qu'à l'emploi des séquences collocatives nominales. Nous avons automatisé l'utilisation de chaque outil de mesure à l'aide d'un programme Python. Certains outils ont déjà été évoqués lors des observations préliminaires (voir partie 2.3) ainsi que lors de l'étape de sélection des modèles (Cf. partie 4.3.3).

5.1 Caractéristiques générales des phrases

Nous allons observer les caractéristiques générales des productions de nos jeux de données afin d'obtenir un aperçu des comportements des sujets humains et des LLMs face à une tâche de production de phrases avec mots imposés. Les résultats nous permettront également de tester nos hypothèses concernant la longueur des phrases et leur diversité lexicale.

5.1.1 Longueur des phrases

Nous avons mesuré la longueur moyenne des productions en calculant le nombre moyen de mots par phrase. Notre unité de mot est délimitée par l'expression régulière "\b\w+\b". La figure 4 compare la longueur des productions chez l'humain et chez les LLMs.

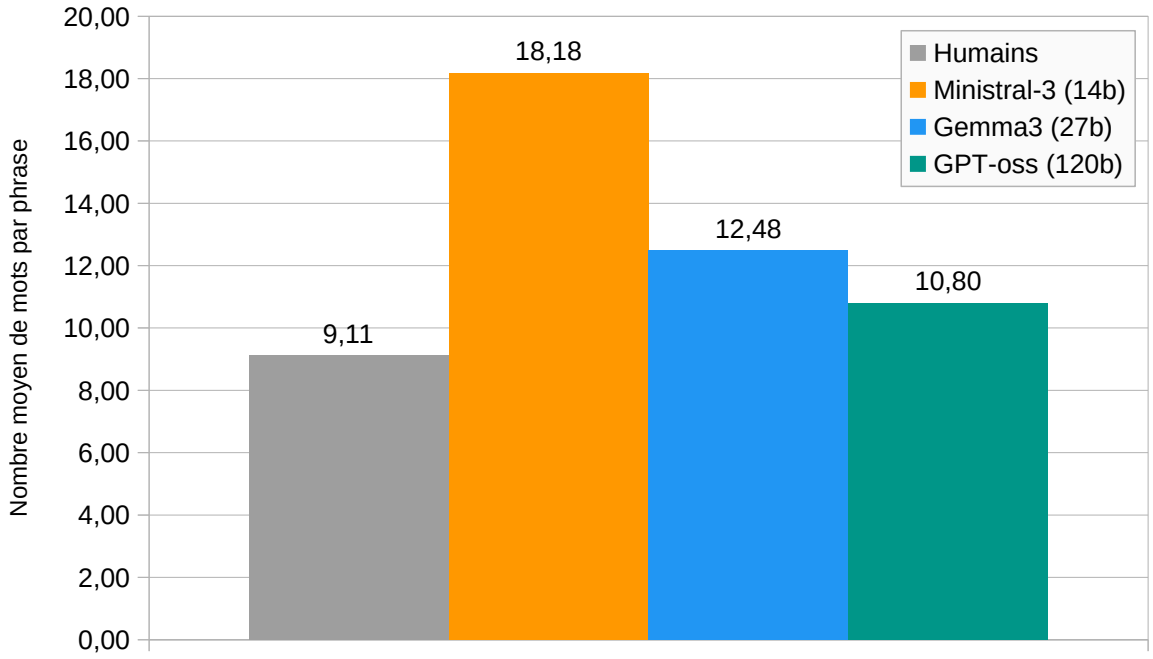

Figure 4 : Nombre moyen de mots par phrase dans les données sollicitées et artificielles

On peut noter que, conformément aux observations d'Alavoine et al. (2024), les LLMs produisent des phrases globalement plus longues que les humains. Ministral-3 se démarque avec des textes près de deux fois plus longs. Le contexte de la tâche semble pouvoir expliquer cet effet de longueur. En effet, l'exercice de rédaction d'une phrase en insérant deux mots imposés est simple à réaliser pour des locuteurs natifs. La faible exigence de la tâche appelle donc une production courte par nature. Nous observons effectivement cette information sur le nombre moyen de mots par phrase sollicitée. Cependant, les LLMs étant de simples modèles statistiques, ils n'ont bien sûr aucune perception des implications inhérentes à la tâche. En rappelant qu'aucune mention de la longueur de la production n'était fournie dans le prompt, il est cohérent de relever un nombre de mots globalement plus élevé dans les phrases des LLMs. Le LLM dont la longueur des productions se rapproche le plus des phrases sollicitées est GPT-oss.

Notre hypothèse se confirme : on observe effectivement des comportements de longueur différents chez les humains et chez l'IA, particulièrement entre Ministral-3 et les sujets humains (longueur doublée pour le modèle). Les LLMs ont, en effet, tendance à produire des phrases plus longues que les humains.

5.1.2 Fréquence et diversité lexicale

Nous avons également souhaité fournir un aperçu de la fréquence lexicale dans les textes sollicités et générés en affichant les 50 mots pleins les plus fréquents sous leur forme lemmatisée. La table de fréquence lexicale est consultable en annexe 1.

Nous nous sommes inspiré de la démarche de Muñoz-Ortiz et al. (2024) pour mesurer la diversité lexicale des données sollicitées et artificielles. Nous avons ainsi utilisé le *Standardized Type-Token Ratio* (STTR) sur des fenêtres de 50 mots parmi la totalité des tokens de chaque jeu de phrases, ainsi que la *Measure of Textual Lexical Diversity* (MTLD). Nous avons employé ces indicateurs grâce à la bibliothèque Python *lexical_diversity* (Kyle, 2018) en les appliquant sur des tokens lemmatisés par *spaCy* (Honnibal et al., 2020) afin de mesurer plus précisément la diversité lexicale des textes. Le tableau 8 présente les valeurs du STTR et de la MTLD des productions analysées. Plus les valeurs sont hautes, plus elles indiquent une richesse lexicale élevée.

	Humains	**Ministral-3 (14b)**	**Gemma3 (27b)**	**GPT-oss (120b)**
STTR	0,53	0,59	0,42	0,43
MTLD	20,67	30,40	19,75	17,40

Tableau 8 : STTR et MTLD appliqués aux phrases sollicitées et artificielles

Les scores de STTR et de MTLD, plus faibles chez Gemma3 et GPT-oss que chez les humains, mettent en avant une plus faible diversité lexicale chez ces modèles. Cependant, les scores de Ministral-3 sont légèrement plus élevés que ceux des humains, remettant en question notre hypothèse selon laquelle les humains utiliseraient un vocabulaire plus varié que les LLMs dans leurs productions.

Naturellement, l'affichage de la fréquence lexicale et la mesure de diversité lexicale reflètent ici une représentation biaisée des comportements rédactionnel et génératif réels des humains et des LLMs due à la tâche d'insertion de mots imposés. On retrouve en effet ces mots parmi les mieux classés dans la table de fréquence lexicale (voir annexe 1), notamment les mots stimuli Evolex qui trouvaient leur place dans un grand nombre de paires (ex : le mot *sac* est associé à 8 paires sur 30 sélectionnées pour nos jeux de données). Ces mesures sont également biaisées par la source des productions : il est délicat de comparer la diversité lexicale d'un jeu de

phrases générées par un unique auteur (phrases artificielles) avec celle d'un jeu de phrases produites par une multitude de locuteurs (phrases sollicitées).

5.2 Caractéristiques inhérentes à la tâche

Nous souhaitons à présent observer deux caractéristiques inhérentes à la tâche de production de phrases avec mots imposés : l'ordre et la flexion des mots insérés dans les phrases par rapport à leur présentation dans les consignes de la tâche (questionnaires et prompt). Afin de mesurer ces critères relatifs aux mots imposés, nous avons d'abord vérifié automatiquement la présence de ces mots dans les phrases de nos jeux de données. A noter que 1141 sur 1166 phrases sollicitées recensent bien les deux mots imposés (97,86%), ainsi que 1493 phrases sur 1500 chez Ministral-3 (99,53%), 1500 sur 1500 chez Gemma3 (100%), et 1499 sur 1500 chez GPT-oss (99,93%). Les taux mesurés par la suite seront basés non sur le total de phrases présentes dans chaque jeu de données, mais sur le total de phrases comprenant les deux mots imposés. Conformément à nos hypothèses, nous nous attendons à observer une plus large conservation des éléments d'ordre et de flexion des mots imposés dans les productions artificielles que dans les phrases sollicitées.

5.2.1 Ordre des mots imposés

Les mots imposés étaient présentés aux participants dans l'ordre [Stimulus – Réponse] Evolex. Nous avons mesuré l'ordre des mots insérés dans les phrases par rapport à leur ordre de présentation dans les consignes de la tâche. On calcule le taux de phrases dont l'ordre des mots imposés est identique à celui présenté dans les consignes parmi les phrases qui contiennent bien les deux mots. Les taux de conservation de l'ordre des mots imposés sont ainsi consultables dans le tableau 9.

	Humains	**Ministral-3 (14b)**	**Gemma3 (27b)**	**GPT-oss (120b)**
ORDRE	82,47 %	69,93 %	75,87 %	83,12 %

Tableau 9 : Taux de conservation de l'ordre des mots imposés par rapport aux consignes

On repère ici que plus de la moitié des phrases artificielles conservent l'ordre des mots imposés par rapport à leur présentation dans le prompt. Aussi, les humains produisent fréquemment les mots dans l'ordre [Stimulus – Réponse] Evolex. Ces taux semblent aller à l'encontre de notre hypothèse selon laquelle les LLMs reproduiraient plus fréquemment l'ordre des mots imposés par le prompt que les humains, plus flexibles dans cette tâche de production.

La démarche représente cependant un biais important : nous avons conservé uniquement les phrases avec insertion de mots en relation de collocation, et 29 paires sur 30 présentaient les mots dans l'ordre des séquences nominales prototypiques qui leur sont associées (ex : pour la paire *Sac – Dos*, les mots sont présentés dans le même ordre que pour la séquence prototypique « sac à dos »). La seule paire qui fait exception est *Noce – Nuit* pour « nuit de noces ». Ainsi, si les humains reproduisent davantage les séquences collocatives comme nous en avons fait l'hypothèse, cela peut expliquer le taux très élevé de conservation de l'ordre des mots par les humains (82,47% ici, contre 59,20% lors de nos observations préliminaires sur les 1659 phrases mélangeant toutes les relations lexico-sémantiques d'Evolex). Cette mesure d'ordre n'est donc pas très pertinente depuis notre recentrage de données. Ce critère reste cependant intéressant à étudier sur des jeux de phrases moins biaisés pour observer si les LLMs prennent moins de liberté et se détachent moins du prompt strict que les sujets humains avec les consignes.

5.2.2 Flexion des mots imposés

Les mots imposés ayant été présentés au singulier dans les consignes de la tâche, nous avons mesuré la flexion des mots par repérage du nombre grammatical des mots en question dans les productions grâce à la bibliothèque Python *spaCy* (Honnibal et al., 2020). On calcule le taux de phrases contenant les deux mots imposés au singulier parmi les phrases qui incluent bien les deux mots. Les taux de conservation de la flexion des mots imposés (au singulier) sont présentés dans le tableau 10.

	Humains	**Ministral-3 (14b)**	**Gemma3 (27b)**	**GPT-oss (120b)**
FLEXION	55,30 %	56,33 %	60,80 %	83,46 %

Tableau 10 : Taux de conservation de la flexion des mots imposés par rapport aux consignes

D'après ces taux, les humains conservent la flexion des mots dans un peu plus de la moitié de leurs productions. GPT-oss s'éloigne des humains, générant presque systématiquement les mots imposés tels qu'il les reçoit dans le prompt (8/10 phrases). Le LLM dont le comportement flexionnel des mots se rapproche le plus des humains est Ministral-3.

Ces taux – globalement plus élevés dans les phrases artificielles que dans les phrases sollicitées – valident notre hypothèse stipulant que les LLMs ont tendance à proposer les mots imposés tels que présentés dans le prompt, tandis que les humains prennent davantage de liberté flexionnelle face aux mots à insérer.

Il est cependant nécessaire de traiter certains cas de flexion à part. En effet, il convient d'analyser séparément les mots qui peuvent se retrouver préférentiellement au pluriel (ex : il est plus fréquent de parler de plusieurs *frites* que d'une seule). C'est le cas pour les mots suivants : *clef/clé* (*Clef – Voiture*), *frite* (*Entrecôte – Frite*), *dame* (*Jeu – Dame*), *noce* (*Noce – Argent* ; *Noce – Nuit* ; *Noce – Or*), *attraction* (*Parc – Attraction*), *huître* (*Parc – Huître*), *bille* (*Sac – Bille*), *commission* (*Sac – Commission*), *spaghetti* (*Spaghetti – Bolognaise* ; *Spaghetti – Carbonara*). Aussi, certains mots se trouvent préférentiellement au pluriel dans des séquences collocatives données (ex : la paire *Parc – Attraction* donne lieu à la séquence « parc d'attractions » : le mot *attractions* est donc rédigé au pluriel bien qu'il soit tout à fait possible et acceptable de le trouver au singulier dans un énoncé). A noter que le lemme *spaghetti*, emprunté à l'italien, est accepté sous les formes *spaghetti* et *spaghettis* au pluriel. Le tableau 11 affiche les taux de phrases présentant une conservation des deux mots imposés au singulier parmi celles qui contiennent bien les deux mots de la paire.

PAIRE DE MOTS	Humains	Ministral-3 (14b)	Gemma3 (27b)	GPT-oss (120b)
Ballon – Rugby	98,00 %	70,00 %	70,00 %	98,00 %
Ballon – Volley	85,29 %	64,00 %	0,00 %	100,00 %
Brochette – Poulet	48,00 %	2,00 %	36,00 %	78,00 %
Clé – Sol	83,33 %	97,96 %	100,00 %	100,00 %
Clef – Voiture	30,00 %	80,00 %	24,00 %	100,00 %
Entrecôte – Frite	13,64 %	2,00 %	4,00 %	40,00 %
Escalier – Marbre	69,39 %	98,00 %	100,00 %	100,00 %
Foie – Canard	50,00 %	40,00 %	100,00 %	80,00 %
Jeu – Dame	65,62 %	24,49 %	98,00 %	84,00 %
Noce – Argent	20,41 %	96,00 %	76,00 %	95,92 %
Noce – Nuit	58,97 %	71,43 %	100,00 %	96,00 %
Noce – Or	7,69 %	36,73 %	22,00 %	48,00 %
Nœud – Cravate	14,29 %	6,00 %	4,00 %	28,00 %
Nœud – Papillon	21,88 %	86,00 %	98,00 %	100,00 %
Parc – Attraction	55,10 %	0,00 %	60,00 %	74,00 %
Parc – Huître	4,76 %	2,00 %	6,00 %	78,00 %
Sac – Bille	10,00 %	50,00 %	82,00 %	92,00 %
Sac – Commission	40,82 %	92,00 %	64,00 %	100,00 %
Sac – Cuir	95,65 %	88,00 %	52,00 %	100,00 %
Sac – Dos	100,00 %	94,00 %	54,00 %	98,00 %
Sac – Luxe	64,00 %	67,35 %	98,00 %	98,00 %
Sac – Main	96,00 %	92,00 %	48,00 %	100,00 %
Sac – Sport	100,00 %	69,39 %	94,00 %	96,00 %
Sac – Voyage	97,87 %	76,00 %	62,00 %	98,00 %
Spaghetti – Bolognaise	4,17 %	26,00 %	30,00 %	44,00 %
Spaghetti – Carbonara	4,26 %	16,00 %	84,00 %	32,00 %
Stylo – Plume	65,96 %	50,00 %	58,00 %	94,00 %
Sucre – Poudre	100,00 %	86,00 %	100,00 %	100,00 %
Tomate – Mozzarella	65,38 %	20,00 %	0,00 %	64,00 %
Trou – Souris	96,15 %	87,76 %	100,00 %	88,00 %

Tableau 11 : Taux de phrases présentant une conservation des deux mots imposés au singulier

Sont représentés en vert tous les taux ≥ 50% ; en rouge < 50%

Ces résultats montrent notamment que les participants humains tendent à adapter la forme grammaticale des mots à l'usage le plus courant associé à la collocation réalisée. Par exemple, pour la paire *Clé – Sol*, les humains davantage les deux mots au singulier, avec des énoncés comme « La partition est en clé de sol. » (humains). Cependant, certaines paires impliquent l'emploi de séquences plus flexibles au niveau du nombre grammatical. Par exemple, pour la paire *Clef – Voiture*, les humains et Gemma3 produisent majoritairement des énoncés du type « J'ai perdu les clefs de la voiture. » (humains), tandis que Ministral-3 et GPT-oss privilégient plus souvent le singulier dans des productions telles que « J'ai perdu la clef de la voiture. » (GPT-oss). Certaines séquences collocatives orientent donc le nombre des noms qui les composent, altérant naturellement les résultats de conservation de la flexion, tandis que d'autres séquences collocatives présentent une plus grande flexibilité quant au nombre grammatical des noms qui les composent.

5.3 Emploi des séquences collocatives nominales

Afin de répondre à notre hypothèse centrale sur l'insertion de mots en relation de collocation, nous avons mesuré l'emploi des séquences collocatives nominales délimitées précédemment (voir partie 4.1) dans les phrases sollicitées et artificielles. Nous avons conçu un programme Python capable de détecter les séquences collocatives par expressions régulières. A noter que notre observation manuelle des phrases sollicitées avait mis en évidence la présence d'un certain nombre d'erreurs orthographiques ou de saisie (ex : « spaguettis » au lieu de « spaghettis » ; « mozzarellla » pour « mozzarella »), d'omissions de diacritiques (ex : « sac a main » au lieu de « sac à main » ; « huitres » pour « huîtres ») et d'erreurs de segmentation textuelle (ex : « dattractions » à la place de « d'attractions » ; « sacà dos » au lieu de « sac à dos »). Ces faits nous ont conduit à baser la détection automatisée des séquences sur des expressions régulières assez lâches dans le but de mesurer avec précision l'emploi réel des séquences collocatives. Ainsi, le programme considère chacune des phrases affiliées à une paire de mots et y recherche la présence des séquences collocatives nominales associées à la paire. A l'issue de ce repérage, nous mesurons le taux global de réalisation des séquences attendues, puis le détail d'apparition de chaque séquence dans les phrases parmi celles qui contiennent bien les deux mots imposés.

L'objectif est double. D'une part, nous souhaitons observer les capacités des LLMs à produire les séquences collocatives attendues à partir des deux noms qui les composent relativement aux humains. D'autre part, nous cherchons à expliquer les tendances observées pour ces emplois.

5.3.1 Aperçu global de la réalisation des séquences

Nous allons commencer par fournir un aperçu global de l'emploi des séquences collocatives nominales attendues dans les phrases sollicitées et artificielles. Nous présentons le taux de réalisation de ces séquences par les humains et par chacun des trois LLMs comparés. Concrètement, il s'agit du nombre de phrases dans lesquelles l'une des séquences collocatives attendues est repérée (parmi les 49 listées en 4.1.2), rapporté au nombre total de phrases contenant bien les deux mots imposés. Ces taux sont présentés en figure 5.

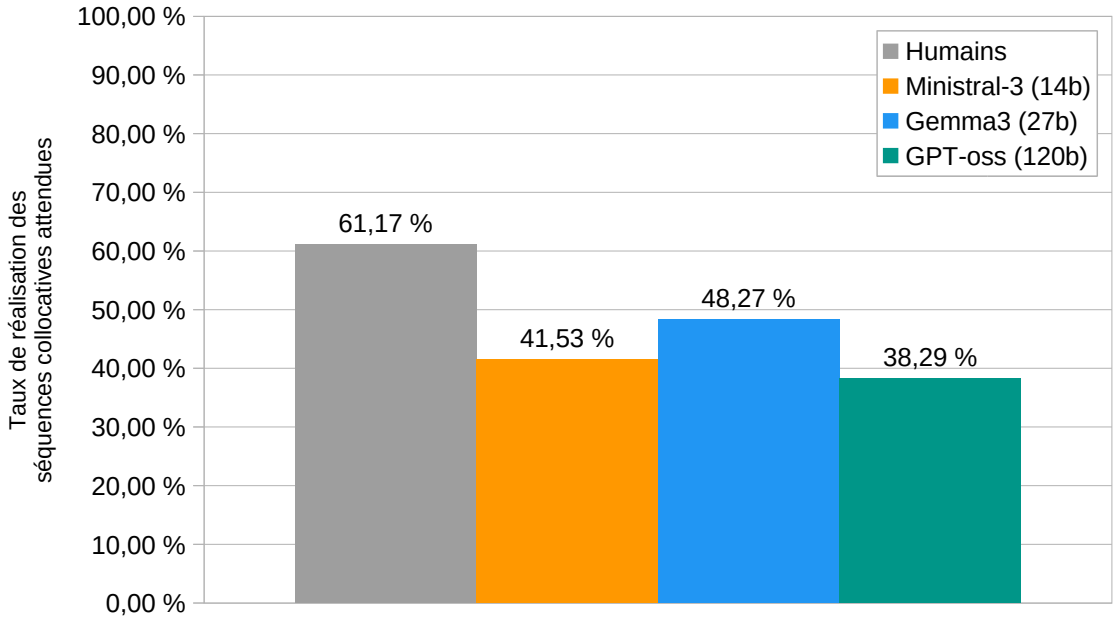

Figure 5 : Taux de réalisation des séquences collocatives nominales attendues

Selon ce graphique, 61,17% des phrases sollicitées contenant les deux mots imposés présentent l'une des séquences collocatives attendues. Ce résultat, globalement élevé, révèle que les humains produisent fréquemment des séquences collocatives à partir de mots isolés entretenant une relation de collocation. Aussi, ce taux représente près de 23 points de pourcentage de plus que chez GPT-oss, presque 20 points de plus que chez Ministral-3, et environ 13 points de plus que chez Gemma3. Le LLM produisant le plus de séquences collocatives attendues est Gemma3. Son emploi des mots imposés en relation de collocation se révèle donc plus proche de celui des humains que les autres modèles comparés. Aussi, le LLM réalisant le moins de séquences collocatives est GPT-oss. Ces taux confirment notre hypothèse selon laquelle, en tâche de production de phrases avec mots imposés en relation de collocation, les humains produisent plus fréquemment des séquences collocatives nominales associées que les LLMs.

5.3.2 Emploi des séquences prototypiques, des variantes et des cas particuliers

Pour préciser ces résultats généraux, nous souhaitons observer le taux d'emploi de chaque séquence collocative nominale dans les phrases sollicitées et artificielles. Pour ce faire, nous calculons la fréquence d'apparition de chaque séquence dans les phrases, rapporté au nombre total de phrases contenant bien les deux mots de la paire. Les taux d'emploi moyen des séquences prototypiques, de leurs variantes et des cas particuliers (voir catégorisation des séquences partie 4.1.3) sont présentés dans le tableau 12. Comme évoqué précédemment, nous représentons et analysons séparément les variantes des séquences prototypiques puisque ces séquences sont mutuellement concurrentes. En effet, les séquences prototypiques reflètent les usages majoritaires en corpus puisque nous avons basé leur catégorisation sous cette dénomination sur les critères conjoints de fréquence en corpus et de structure syntaxique (Cf. parties 4.1.2 et 4.1.3). A l'inverse, les variantes sont les séquences moins fréquentes associées aux mêmes paires de mots. Aussi, les cas particuliers (« entrecôte avec des frites » et « foie gras de canard ») sont présentés dans cette vue d'ensemble et seront discutés plus en détail ultérieurement.

CATÉGORIE DES SÉQUENCES	Humains	Ministral-3 (14b)	Gemma3 (27b)	GPT-oss (120b)
Séquences prototypiques	47,24 %	37,00 %	37,32 %	29,51 %
Variantes	11,64 %	21,92 %	14,00 %	21,47 %
Cas particuliers	37,39 %	19,00 %	61,00 %	0,00 %

Tableau 12 : Taux moyens d'emploi des séquences prototypiques, des variantes et des cas particuliers

Globalement, 47,24% des phrases sollicitées contiennent une séquence collocative prototypique. Cela représente presque 18 points de plus que chez GPT-oss, environ 10 points de plus que chez Ministral-3 et Gemma3. Les deux cas particuliers sont très largement employés par Gemma3, et en moindre mesure par les humains. Nous observons également des différences notables entre l'emploi des séquences prototypiques et les variantes chez les différents participants (humains et LLMs). Nous allons nous appuyer sur une représentation graphique afin de mieux analyser cette observation.

Rappelons que les séquences prototypiques reflètent les usages majoritaires en corpus. Ainsi, la comparaison du taux moyen d'emploi des séquences prototypiques avec celui des variantes permettrait de mettre en avant une éventuelle préférence de l'un des deux types de séquences par les humains et les LLMs dans les phrases sollicitées et artificielles. La figure 6 donne une représentation graphique de cette comparaison.

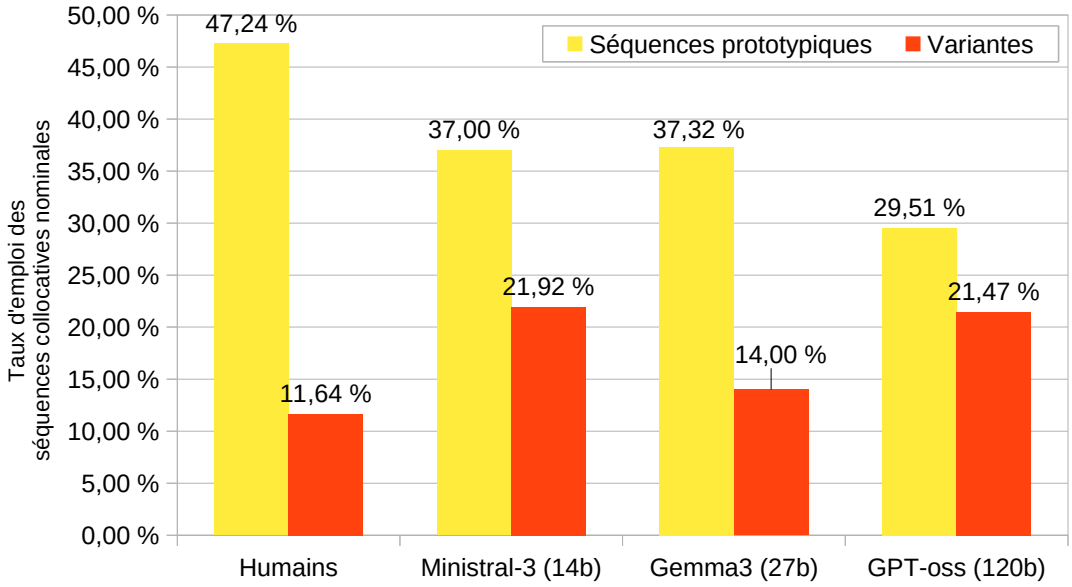

Figure 6 : Taux moyens d'emploi des séquences prototypiques et des variantes

On repère ici une nette préférence des humains et des LLMs pour les séquences prototypiques. Cette information met en évidence une convergence entre les emplois préférés dans les phrases de nos jeux de données et les usages attestés en corpus. En effet, les LLMs – dont l'apprentissage est théoriquement basé sur des données de même nature que celles du corpus – et les humains présentent une préférence pour l'emploi des séquences collocatives les plus présentes en corpus (séquences prototypiques) par opposition aux variantes concurrentes. Aussi, les phrases artificielles recensent davantage d'emplois de variantes concurrentes que les phrases sollicitées avec une différence allant de 2 à 10 points de pourcentage environ.

Ainsi, ces observations suggèrent la tendance des humains et des LLMs à proposer les mêmes comportements qu'en corpus vis-à-vis des séquences prototypiques. Elles montrent aussi que les LLMs utilisent légèrement plus de variantes que les humains.

5.3.3 Détails de l'emploi des séquences prototypiques par structure syntaxique

Nous souhaitons à présent observer précisément l'emploi des séquences prototypiques selon leur type de structure syntaxique afin d'avoir une meilleure vision de l'utilisation de ces séquences dans les phrases produites. Les taux moyens d'emploi des séquences prototypiques selon leur structure syntaxique sont présentés en figure 7.

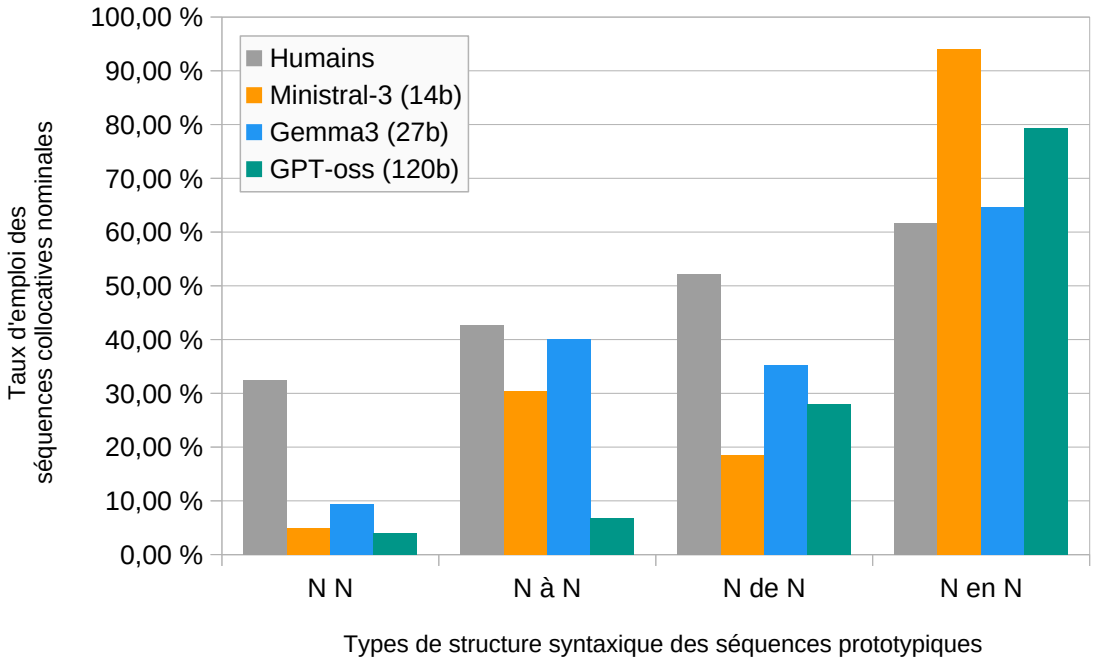

Figure 7 : Taux moyens d'emploi des séquences prototypiques selon leur structure syntaxique

Cette représentation nous permet d'observer une utilisation plus importante des séquences prototypiques en structures **N N** (ex : *nœud papillon*), **N à N** (ex : *parc à huîtres*) ainsi que **N de N** (ex : *nuit de noces*) par les humains que par les LLMs. C'est cependant le contraire pour les séquences en structure **N en N** (ex : *escalier en marbre*). Aussi, proportionnellement, les séquences en **N en N** restent plus largement proposées par tous les participants (humains et LLMs) que les autres types de structures. Il est toutefois essentiel de noter que les effectifs des séquences par type de structure syntaxique sont très hétérogènes : **N N** compte 6 séquences, **N à N** en recense 4, **N de N** en regroupe 19, tandis que **N en N** en compte 3.

5.3.4 Analyse de la saillance des séquences

Puisque nous avons mesuré individuellement le taux d'emploi de chaque séquence collocative nominale considérée dans les phrases analysées, il est possible d'aller observer précisément lesquelles sont privilégiées par les participants (humains et LLMs). Les taux précis des réalisations de chaque séquence par chacun des participants sont consultables en annexe 2. Ces résultats nous ont permis de créer une classification des séquences d'après leur degré de saillance. Le principe de saillance repose ici sur l'apparition plus ou moins fréquente d'une séquence collocative dans les phrases sollicitées et artificielles qui refléterait le degré "d'évidence" de cette séquence pour les humains et les LLMs. Nous avons alors défini des seuils permettant de classifier les séquences collocatives selon si elles sont évidentes pour les participants (taux d'emploi de la séquence supérieur ou égal à 50%) ou non-évidentes (taux inférieur à 50%). Nous avons classifié chacune des paires selon les catégories suivantes :

I. Séquences évidentes pour les humains et les LLMs (H ≥ 50 % & au moins 2 LLMs ≥ 50 %)
II. Séquences évidentes pour les humains uniquement (H ≥ 50 % & maximum 1 LLM ≥ 50 %)
III. Séquences évidentes pour les LLMs uniquement (H < 50 % & au moins 2 LLMs ≥ 50 %)
IV. Séquences évidentes pour aucun d'entre eux (H < 50 % & maximum 1 LLM ≥ 50 %)

Afin d'observer le degré de saillance des séquences pour les humains par opposition aux LLMs, nous avons décidé de regrouper les trois LLMs – Ministral-3 (14b), Gemma3 (27b), GPT-oss (120b) – en une unique catégorie ("LLMs"). Les tableaux 13-A et 13-B permettent de visualiser la classification des séquences d'après leur degré de saillance auprès des humains et des LLMs.

CATÉGORIE	SÉQUENCE	Humains	LLMs
I Séquences évidentes pour tous	ballon de rugby brochette de poulet parc d'attractions sac en cuir sac à dos sac à main sac de sport sucre en poudre	✓	✓
II Séquences évidentes pour les humains uniquement	clé de sol / clef de sol jeu de dames noces d'argent nuit de noces noces d'or nœud papillon parc à huîtres sac de billes sac de luxe trou de souris	✓	✗
III Séquences évidentes pour les LLMs uniquement	escalier en marbre nœud de ma/ta/…/leur cravate spaghettis à la bolognaise spaghettis à la carbonara	✗	✓

Tableau 13-A : Catégorisation des séquences selon leur degré de saillance

CATÉGORIE	SÉQUENCE	Humains	LLMs
IV Séquences évidentes pour aucun	ballon de volley clé/clef de voiture clé/clef de la voiture clé/clef de ma/ta/…/leur voiture entrecôte avec des frites entrecôte frites escalier de marbre foie gras de canard foie de canard jeu des dames nuit des noces nuit de mes/tes/…/leurs noces nœud de cravate nœud de la cravate nœud de papillon sac à billes sac à commissions sac de commissions sac de cuir sac de voyage spaghettis bolognaise spaghettis carbonara stylo plume stylo à plume sucre poudre tomate mozzarella tomate à la mozzarella	✗	✗

Tableau 13-B : Catégorisation des séquences selon leur degré de saillance

Cette représentation donne ainsi une vision claire du fait suivant : sur les 49 séquences collocatives considérées, plus de la moitié sont peu voire pas produites par les humains et les LLMs (catégorie IV).

Nous remarquons également que toutes les séquences évidentes pour les humains (catégories I et II) sont en fait identiques à celles que nous avons désignées comme prototypiques suite à nos observations en corpus (Cf. partie 4.1.3). Ainsi, la plupart des séquences des catégories I et II font partie des formes les plus fréquentes en corpus proportionnellement à toutes les séquences étudiées. C'est par exemple le cas pour « sac à dos », « sac à main », « parc d'attractions » ou encore « sucre en poudre » (fréquences en corpus présentées dans la partie 4.1.2). A l'inverse, les séquences évidentes pour les LLMs et peu employées par les humains (catégorie III) comptent 3 variantes concurrentes minoritaires (donc plus rares en corpus) et une seule prototypique.

Aussi, les LLMs semblent résister à produire les séquences observables en catégorie II (e.g « clé de sol », « nœud papillon », « parc à huîtres ») par rapport aux séquences en catégorie I (ex : « parc d'attractions », « sac à dos », « sucre en poudre »). Étant donné que les LLMs apprennent en partie sur des données de corpus, cette résistance peut s'expliquer par l'influence des tendances dans ces corpus, qui visent eux-mêmes à représenter les usages linguistiques des humains. En effet, nous pensons par exemple que l'expression « parc d'attractions » (catégorie I) sera plus fréquente que « parc à huîtres » (catégorie II) dans les usages réels. Il en va de même pour « sac à dos » ou « sac à main » (catégorie I) qui semblent être des expressions typiques du quotidien, à la différence de « jeu de dames » ou « noces d'or » (catégorie II) qu'on suppose être utilisées plus occasionnellement. Ces assertions sont confirmées par les fréquences en corpus et pourraient expliquer la différence d'emploi des séquences des catégories I et II par les LLMs. Néanmoins, l'effet de fréquence en corpus n'est pas le seul facteur en cause dans la résistance des LLMs à produire ces formes. En effet, les séquences présentes en catégorie II semblent être très figées et non-compositionnelles en comparaison avec les séquences de catégorie I qui relèvent de structures plus souples et qui semblent davantage compositionnelles (Polguère, 2003). Pour exemple, « noces d'or » et « nœud papillon » (catégorie II) sont des expressions linguistiquement peu transparentes pour renvoyer au concept ou au référent qu'elles désignent. Par opposition, le sens des séquences « brochette de poulet » et « sac en cuir » (catégorie I) peut être compris aisément par la somme du sens de leurs unités (principe de compositionnalité). Ces phénomènes de figement et de compositionnalité pourraient donc expliquer le sous-emploi de certaines séquences collocatives nominales (ici classées en catégorie II) par les LLMs.

Enfin, parmi les séquences non-évidentes pour les humains et les LLMs (catégorie IV), on observe un grand nombre de variantes concurrentes minoritaires (ex : « nœud de papillon », « jeu des dames », « nuit des noces »). La préférence pour leur forme homologue (majoritaire en corpus) dans les phrases sollicitées et artificielles vient expliquer le sous-emploi de ces séquences. Néanmoins, on observe aussi une faible saillance pour des séquences majoritaires en corpus qu'on considérait comme prototypiques et qu'on estimait fréquentes et plutôt figées, notamment « stylo plume », mais également « spaghettis bolognaise », « spaghettis carbonara » et « tomate mozzarella ». Ces résultats nous montrent donc que ces séquences ne sont pas si évidentes, même pour les humains qui ont parfois alterné l'emploi de séquences concurrentes de façon équivalente. En effet, les emplois collocatifs sont partagés entre « spaghettis bolognaise » et « spaghettis à la bolognaise », ou encore « spaghettis carbonara » et « spaghetti à la carbonara ». Autrement, les humains ont parfois préféré massivement des structures alternatives comme « tomate et la mozzarella » ou « tomate et de la mozzarella » qui relèvent ici de la coordination et non d'une structure collocative nominale tel que nous l'avons définie dans ce travail.

Les cas particuliers « entrecôte avec des frites » et « foie gras de canard » ainsi que leur séquence homologue minoritaire en corpus (« entrecôte frites » et « foie de canard ») sont tous peu représentés dans les phrases sollicitées et artificielles, comme l'indique leur présence commune dans la catégorie IV. Conformément aux fréquences en corpus, les formes « entrecôte avec des frites » et « foie gras de canard » restent toutefois préférées par les humains et par les LLMs par rapport aux séquences homologues. On note des variantes récurrentes chez les LLMs comme chez les humains : « entrecôte et des frites » (humains, Gemma3), « entrecôte accompagnée de frites » (Ministral-3, GPT-oss) ; « foie du canard » (GPT-oss).

5.3.5 Alternatives aux séquences attendues

Nous souhaitons observer certains cas de sous-emploi des séquences collocatives attendues dans les phrases analysées. Précisément, nous notons les séquences dont la classification des tableaux 13-A et 13-B retient particulièrement notre attention par rapport aux observations menées en corpus et à notre perception de leur degré de figement. Nous observons les configurations alternatives proposées dans les phrases sollicitées et artificielles afin de présenter une liste de phrases-exemples contenant des cas fréquents de structures alternatives. Le tableau 14 présente des exemples de formulations alternatives lorsque les séquences attendues n'apparaissent pas dans les phrases sollicitées et artificielles.

	Humains	**LLMs**
Séquence attendue	**Exemples de phrases avec structures alternatives**	**Exemples de phrases avec structures alternatives**
clé/clef de sol	–	« La clé du sol a rouillé sous la pluie. » (Gemma3) « La clé de cette énigme réside dans le sol ancien où dorment les secrets oubliés. » (Ministral-3) « La clé du piano repose sur le sol. » (GPT-oss)
clé/clef de voiture	–	« La clé de ma nouvelle voiture est accrochée à un porte-clés en forme de licorne. » (Ministral-3)
escalier de/en marbre	« L'escalier est en marbre. » « L'escalier est fait de marbre. »	–
jeu de dames	–	« Le jeu de la dame révèle une stratégie subtile. » (GPT-oss) « La dame a perdu le jeu aux échecs. » (GPT-oss)
nuit de noces	–	« La nuit de la noce, les mariés étaient radieux. » (Gemma3)
noces d'or	–	« L'or scintillait sur la noce du village. » (GPT-oss)
nœud papillon	–	« Il a passé la soirée à ajuster le nœud de son papillon en soie. » (Gemma3) « Le papillon se posa délicatement sur le nœud de la cravate en soie. » (Ministral-3)
parc à huîtres	–	« Dans le parc, j'ai dégusté une huître fraîche au bord du lac. » (GPT-oss)
sac de billes	–	« Elle a trouvé une bille au fond de son sac. » (Gemma3)
sac à/de commissions	« J'ai pris mon sac pour faire des commissions. »	« Après avoir fait ses courses, elle a déposé le sac à la commission de contrôle. » (Gemma3)
sac de voyage	« Je prépare mon sac pour partir en voyage. »	« J'ai préparé mon sac pour un long voyage. » (Gemma3)
spaghettis (à la) bolognaise	« Je n'aime pas les spaghettis en sauce bolognaise. »	–
stylo plume	« La plume et le stylo sont des moyens pour écrire. »	« Avant de choisir entre son stylo à bille et ses vieilles plumes d'oie, il hésita longuement pour écrire ce poème. » (Ministral-3b)
tomate mozzarella	« J'aime les pizzas à la tomate et à la mozzarella. »	« J'ai préparé une salade de tomates et mozzarella pour le déjeuner. » (GPT-oss)

Tableau 14 : Exemples de formulations alternatives aux séquences attendues

Ainsi, ces exemples mettent en avant plusieurs cas. D'abord, plusieurs structures observées indiquent que la relation entre certains mots imposés n'est pas uniquement syntagmatique : les mots relèvent du phénomène de relations multiples (voir partie 1.1.4 sur les relations lexicales multiples). On retrouve par exemple, pour la paire de mot *Escalier* – Marbre, la phrase « L'escalier est fait de marbre. » qui signale une relation de méronymie entre les deux noms. Les phrases-exemples contenant les paires *Stylo* – *Plume* et *Tomate* – *Mozzarella* mettent en évidence la relation de co-hyponymie qui relie ces lexies. On repère alors une préférence pour la coordination entre les deux mots imposés lorsqu'ils sont co-hyponymes (ex : « la plume et le stylo » ; « tomates et mozzarella »). Enfin, les phrases-exemples soulignent la dominance des relations associatives pour certaines paires de mots telles que *Spaghetti* – *Bolognaise* avec des structures telles que « spaghettis en sauce bolognaise » qui témoignent effectivement de l'association d'idées entre les deux concepts désignés par les lexies (*spaghettis* = pâtes et *bolognaise* = sauce). On retrouve ce même phénomène avec la paire *Sac* – *Voyage* pour laquelle la relation associative semble prévaloir sur la relation syntagmatique (ex : des structures comme « sac pour partir en voyage » sont privilégiées par rapport à « sac de voyage »).

Aussi, nous observons ici des séquences qui s'apparentent à des variantes structurelles des formes prototypiques, mais qui désignent en réalité d'autres référents. Nous ne les avions pas retenues dans notre liste de variantes concurrentes puisque leur fréquence en corpus n'indiquait aucune dominance particulière de ces formes. Nous retrouvons alors les séquences « jeu de la dame » et « nuit de la noce » dans les productions des LLMs.

Nous observons seulement une alternative se rapprochant du schéma des séquences collocatives nominales prototypiques : il s'agit de « clé du sol ». Cette séquence a été produite 14 fois par Gemma3 (sur 50 phrases avec la paire *Clé* – *Sol*), et une fois par GPT-oss dans la phrase « La clé du sol ouvre la porte de la musique. »). Cette séquence semble être l'objet d'une reproduction erronée de la séquence collocative « clé de sol » par les LLMs.

Les autres structures observées dans les phrases témoignent d'une préférence pour l'insertion de multiples lexies entre les mots imposés. Cela a pour conséquence d'éloigner plus ou moins considérablement les deux noms cibles. Dans les phrases-exemples observées chez les humains, l'éloignement des deux noms suggère plutôt un degré de figement peu élevé des séquences considérées. Toutefois, chez les LLMs, l'éloignement des mots cibles – en particulier lorsque l'on attend des formes telles que « jeu de dames », « noces d'or », « nœud papillon » ou « parc à huîtres » – semble témoigner d'une réelle faiblesse des systèmes d'IA à produire certaines séquences lexicalisées plutôt figées du langage naturel.

Conclusions et perspectives

Ce travail visait à comparer des phrases produites par l'humain et générées par système d'IA en tâche de production de phrases avec mots imposés. Le but était d'abord d'en noter les différences linguistiques générales (structurales et lexicales) ainsi que celles inhérentes à la tâche. L'objectif central de l'étude résidait dans l'analyse des capacités des LLMs à produire des séquences complexes du langage naturel en vue de proposer des explications aux tendances observées.

Afin de conduire nos analyses comparatives, nous avons constitué deux jeux de phrases produites à partir de deux noms imposés : le jeu de phrases sollicitées, regroupant des énoncés rédigés par des humains, et le jeu de phrases artificielles, rassemblant des productions générées par trois LLMs différents (Ministral-3 (14b), Gemma3 (27b), GPT-oss (120b)).

Nous avons testé chacune de nos hypothèses en automatisant le traitement des phrases grâce à un programme Python et en conduisant une analyse statistique et linguistique des résultats. L'étude comparative a été structurée en trois axes : caractéristiques générales des phrases (structurales et lexicales) correspondant aux hypothèses (1) et (2), caractéristiques inhérentes à la tâche de production de phrases avec mots imposés (analyse morpho-syntaxique) selon les hypothèses (3) et (4), emploi des séquences collocatives nominales à partir des deux noms qui les composent, se rapportant à l'hypothèse (5).

Nous avons d'abord observé des caractéristiques générales des phrases (structurales et lexicales). Pour ce faire, nous avons mesuré la longueur par indice du nombre moyen de mots par production. Les résultats montrent que, conformément à notre hypothèse (1), il existe une différence significative de longueur entre les phrases sollicitées et artificielles : les LLMs comparés produisent des phrases jusqu'à deux fois plus longues en moyenne. Nous avons également exploré les différences lexicales en proposant un tableau des mots pleins les plus fréquents dans les phrases de chaque participant (humains ou LLMs), ainsi qu'en mesurant la diversité lexicale par les indices de STTR et MTLD. Les résultats de ces deux mesures ont indiqué un vocabulaire légèrement plus varié dans les données sollicitées que dans les productions de Gemma3 et GPT-oss. Ces résultats semblaient confirmer notre hypothèse (2). Cependant, le jeu de phrases générées par Ministral-3 a obtenu des scores de diversité lexicale supérieurs à ceux du jeu de phrases sollicitées : les premiers résultats n'étaient donc pas extrapolables à tous les LLMs testés.

Nous avons ensuite exploré les comportements relatifs à l'insertion de mots imposés dans les phrases produites. Concrètement, nous avons vérifié si l'ordre d'apparition des mots dans les consignes de la tâche influençait l'ordre des mots dans les phrases produites. Notre hypothèse (3), stipulant que les LLMs se conformeraient davantage à l'ordre d'apparition des mots dans le prompt que les humains, est invalidée par nos résultats : les phrases sollicitées, dans leur grande majorité, conservaient l'ordre des mots au même titre que les phrases artificielles. Nous avons également observé si la forme exacte des mots imposés dans les consignes conditionnait la forme produite en observant précisément la flexion des noms dans les phrases. Notre hypothèse (4), qui suggérait que les LLMs prendraient moins de liberté dans la flexion des mots imposés, est vérifiée par les résultats : un peu plus de la moitié des phrases sollicitées présentait une conservation du nombre grammatical des mots imposés, contre environ 6/10 productions de Gemma3, et 8/10 phrases générées par GPT-oss. Ainsi, les LLMs – en particulier Gemma3 et GPT-oss ici –

semblaient effectivement se limiter à reproduire les mots tels que traités dans le prompt sans prendre de liberté flexionnelle.

Enfin, si les travaux de Hashiloni et al. (2025) sur les expressions idiomatiques et les MWEs ont permis d'afficher la difficulté des LLMs à identifier ces séquences flexibles, inconstantes et complexes du langage naturel, l'objectif central de ce mémoire consistait à tester la capacité des modèles à produire de tels objets, et plus précisément ce que nous avons appelé les *séquences collocatives nominales*. La comparaison des phrases sollicitées et artificielles a alors permis de tester notre hypothèse (5) selon laquelle, en tâche de génération de phrases avec deux noms imposés entretenant une relation de collocation, les humains produiraient davantage de séquences collocatives associées que les LLMs. Conformément à cette hypothèse, nos résultats démontrent que les LLMs produisent moins fréquemment les séquences collocatives attendues que les humains (6 phrases sollicitées sur 10 proposent l'une des séquences attendues, contre presque 4/10 à 5/10 phrases artificielles). Aussi, une analyse détaillée de l'emploi des séquences selon leur fréquence en corpus et leur structure syntaxique a permis de mettre en évidence une préférence des participants pour des séquences prototypiques (fréquentes en corpus) au détriment de leurs variantes syntaxiques (minoritaires en corpus). La classification de chaque séquence étudiée selon leur saillance dans les phrases sollicitées et artificielles a permis de préciser les tendances de sous-emplois de certains types de séquences par les participants. Ainsi, une analyse statistique et linguistique des résultats nous a amené à proposer des explications de ces tendances. Nous retenons notamment deux considérations. D'abord, nous pensons que les séquences peu fréquentes en corpus impliquent un sous-emploi de ces formes par les LLMs puisque ceux-ci sont entraînés sur des données de corpus. Aussi, nous supposons que la nature figée de certaines séquences collocatives constitue un challenge pour les LLMs, indépendamment de leur fréquence en corpus, en raison de leur caractère complexe et inconstant sur le plan compositionnel.

Notre méthodologie présente certaines limites qu'il convient de discuter. D'abord, les mesures de diversité lexicale se trouvaient limitées par la tâche : la présence de mots imposés dans les productions a naturellement fait chuter les résultats censés rendre compte de la richesse lexicale proposée par les participants. Aussi, l'analyse de l'ordre et de la flexion des mots imposés en réponse aux hypothèses (3) et (4) présentait une incertitude majeure. En effet, il nous était impossible de connaître avec certitude l'ordre dans lequel ont été présentés les mots dans les questionnaires *Evolex* dont les phrases sont issues. Il en va de même pour le nombre grammatical des mots imposés. Ce doute rendait donc l'extrapolation de ces résultats sur les phrases sollicitées peu fiable. Il aurait ainsi été préférable d'avoir un contrôle total sur le protocole expérimental menant à la production des phrases sollicitées. Aussi, la poursuite de l'hypothèse (5), qui nous a conduit à recentrer nos données d'origine sur des phrases produites uniquement à partir de noms en relation de collocation, a restreint notre échantillon d'analyse de manière trop radicale pour nos hypothèses (1-4). Enfin, l'hypothèse (5) sur l'emploi des séquences collocatives semblait interférer avec les hypothèses (3) et (4) concernant l'ordre et la flexion des mots puisqu'un emploi fréquent de séquences collocatives par un participant favorisait d'une part un certain ordre récurrent des mots imposés selon la séquence collocative produite, et d'autre part l'emploi préférentiel d'un nombre grammatical plus adapté à la séquence collocative. L'ordre et la flexion des mots imposés dans les phrases ne seraient donc pas expliqués exclusivement par leur apparition dans les consignes de la tâche, mais également par l'emploi éventuel de séquences collocatives, rendant impossible de tirer des conclusions fiables à partir des hypothèses (3) et (4). Il aurait donc été nécessaire de diviser les analyses en utilisant, dans un premier temps, des jeux

de données plus généraux pour vérifier les hypothèses (1), (2), (3) et (4), avant de s'appuyer sur notre propre jeu de données pour l'hypothèse (5).

Ce travail nous a donc permis de proposer une comparaison de phrases sollicitées et artificielles afin d'en relever certaines différences linguistiques globales, ainsi que de conduire une analyse des comportements humains et artificiels face à une tâche de production de phrases avec mots imposés. Cette étude a également permis de mettre en évidence la fragilité des LLMs à produire des séquences collocatives nominales, notamment lorsqu'elles sont peu fréquentes en corpus, ou fortement figées et non-compositionnelles.

Il serait intéressant de reposer la question de l'emploi des séquences collocatives nominales dans des jeux de phrases sollicitées et artificielles en conduisant une analyse lexicologique précise du degré de figement et de compositionnalité des séquences analysées. Ceci permettrait de confirmer s'il s'agit bien d'un facteur déterminant pour expliquer les contrastes de réalisation de telles séquences par les humains et les LLMs.

Ce mémoire a révélé une relative fragilité de performances des LLMs quant à la production de certaines séquences collocatives. Les explications à cette faiblesse évoquent la nécessité d'implémenter davantage de séquences plus figées, non-compositionnelles et moins fréquentes du français dans des corpus, afin que les LLMs apprennent à les reproduire de manière plus systématique. La maîtrise de cet objet linguistiquement complexe permettrait ainsi aux LLMs de produire des textes d'autant plus proches du langage naturel.

Malgré leur capacité à générer des textes linguistiquement et conceptuellement cohérents approchant du langage humain, les LLMs présentent donc encore des résistances dans l'imitation de certains comportements langagiers. Les systèmes d'IA devraient être optimisés afin d'être en mesure de mieux mobiliser les subtilités du langage naturel en tâche de génération textuelle.

Bibliographie

Alavoine, N., Coavoux, M., Esperança-Rodier, E., Gallienne, R., González-Gallardo, C. E., Goulian, J., Moreno, J. G., Névéol, A., Schwab, D., Segonne, V., Simoens, J. (2024). Limitations of human identification of automatically generated text. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Italia, (pp. 10511-10516).

Cai, B., Ng, C. B. L., Tan, D., & Hotama, S. (2024). Low-cost generation and evaluation of dictionary example sentences. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico, (pp. 3538–3549).

Cruse, D. A. (1986). *Lexical semantics*. Cambridge : Cambridge University Press.

De Saussure, F. (1916). Cours de linguistique générale. Paris : Payot.

Díaz, O. M. (2009). Les expressions lexicalisées: Schémas linguistiques. *Lenguas Modernas*, (33), 133–152.

Gaume, B., Tanguy, L., Fabre, C., Ho-Dac, L. M., Pierrejean, B., Hathout, N., Farinas, J., Pinquier, J., Danet, L., Péran, P., De Boissezon, X., Jucla, M. (2018). Automatic analysis of word association data from the Evolex psycholinguistic tasks using computational lexical semantic similarity measures. In *13th International Workshop on Natural Language Processing and Cognitive Science (NLPCS),* Poland.

Gemma Team: Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., … & Hussenot, L. (2025). Gemma 3 technical report. (arXiv:2503.19786)

Gross, G. (1988). Degré de figement des noms composés. *Langages*, (90), 57–72.

Gross, G. (1996). *Les expressions figées en français: noms composés et autres locutions.* Éditions Ophrys.

Hashiloni, K. G., Hefetz, O., & Bar, K. (2025). Easy as PIE? identifying multi-word expressions with LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 23782–23801).

Hattouti, J., Gil, S., & Laval, V. (2016). Le développement de la compréhension des expressions idiomatiques: une revue de littérature. *L'Année psychologique, 116*(1), 105–136.

Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.

Jurafsky, D., & Martin, J. H. (2025). *Speech and Language Processing* (3rd ed. draft). Stanford University.

Kyle, K. (2018). *lexical_diversity*. GitHub. (https://github.com/kristopherkyle/lexical_diversity)

Marquer, P. (1994). *La compréhension des expressions idiomatiques. L'année psychologique, 94*(4), 625–656.

Meta AI. (2024). *Llama 3.2 model card.* Hugging Face. (https://huggingface.co/meta-llama/Llama-3.2-3B)

Microsoft Team: Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., … & Zhou, X. (2024). Phi-3 technical report: A highly capable language model locally on your phone. (arXiv:2404.14219)

Microsoft Team: Abouelenin, A., Ashfaq, A., Atkinson, A., Awadalla, H., Bach, N., Bao, J., … & Zhou, X. (2025). Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. (arXiv:2503.01743)

Mistral AI Team: Liu, A. H., Khandelwal, K., Subramanian, S., Jouault, V., … & Ramzi, Z. (2026). Ministral 3. (arXiv:2601.08584)

Morris, J., & Hirst, G. (2004). Non-classical lexical semantic relations. In *Proceedings of the computational lexical semantics workshop at HLT-NAACL 2004* (pp. 46–51).

Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., Awadallah, A. (2023). Orca: Progressive Learning from Complex Explanation Traces of GPT-4. (arXiv:2306.02707)

Muñoz-Ortiz, A., Gómez-Rodríguez, C., & Vilares, D. (2024). Contrasting linguistic patterns in human and LLM-generated news text. *Artificial Intelligence Review, 57*(10), 265.

Ollama. (2026a). *Ollama.* Ollama. (https://ollama.com/)

Ollama. (2026b). *ollama-python.* GitHub. (https://github.com/ollama/ollama-python)

Ollama. (2026c). *Cloud*. Ollama. (https://docs.ollama.com/cloud)

OpenAI (2025a). *Introducing GPT-5.2*. OpenAI. (https://openai.com/index/introducing-gpt-5-2/)

OpenAI (2025b). *Introducing gpt-oss*. OpenAI. (https://openai.com/index/introducing-gpt-oss/)

Polguère, A. (2003). *Lexicologie et sémantique lexicale notions fondamentales*. Presses de l'Université de Montréal.

Qwen Team: Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., … & Zhu, T. (2023). Qwen Technical Report. (arXiv:2309.16609)

Qwen Team: Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., … & Meng, Z. (2024) Qwen2.5 Technical Report. (arXiv:2412.15115)

Rapp, R. (2002). The computation of word associations: comparing syntagmatic and paradigmatic approaches. In *COLING 2002: The 19th International Conference on Computational Linguistics*, Taiwan.

Riegel, M. (1994). *Grammaire Méthodique du Français*. Paris : Presses Universitaires de France.

Rodriguez, M. A., Candito, M., & Huyghe, R. (2025). FAMWA: A new taxonomy for classifying word associations (which humans improve at but LLMs still struggle with). In *Proceedings of the 16th International Conference on Computational Semantics* (pp. 175-188).

Sanacore, Daniele. (2024). *Une histoire de famille : description morphosémantique des lexèmes construits et des relations dérivationnelles*. Thèse de doctorat. Université Toulouse Jean Jaurès, Toulouse.

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., ... & Resnik, P. (2024). *The Prompt Report: A Systematic Survey of Prompt Engineering Techniques*. (arXiv preprint arXiv:2406.06608).

Service Universitaire de Pédagogie (Université Bretagne Sud). (2025). *Comment optimiser ses prompts (prompt engineering) ?* (https://sup-ubs.fr/faq/comment-optimiser-ses-prompts/)

Simounet, Mathilde. (2021). *Prédiction automatique d'associations lexicales : comparaison de données expérimentales et naturelles*. Mémoire de Master 2. Université Toulouse Jean Jaurès, Toulouse.

Solaiman, I. (2023). The gradient of generative AI release: Methods and considerations. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency* (pp. 111-122).

Taher, M. D., & Salih, S. M. (2022). A Paradigmatic Lexical Relation Study of Analysing Entailment. In *Identity and Inclusion Relations. Koya University Journal of Humanities and Social Sciences*, *5*(1), 159-166.

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., ... & Hu, X. (2024). Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. *ACM Transactions on Knowledge Discovery from Data, 18*(6), 1-32.

Annexes

Annexe 1 : Table de fréquence lexicale des phrases sollicitées et artificielles

RANG	Humains Lemmes	Humains Fréquence	Ministral-3 Lemmes	Ministral-3 Fréquence	Gemma3 Lemmes	Gemma3 Fréquence	GPT-oss Lemmes	GPT-oss Fréquence
1	sac	344	sac	406	sac	412	sac	403
2	noce	117	noce	153	noce	156	préparer	172
3	spaghetti	96	vieux	138	jouer	134	noce	150
4	ballon	86	préparer	118	parc	128	acheter	121
5	parc	72	clé	117	préparer	125	parc	103
6	manger	69	cuir	110	enfant	106	ballon	101
7	clé	68	parc	107	soir	104	nœud	100
8	prendre	57	petit	106	nœud	101	spaghetti	100
9	main	57	ballon	102	ballon	100	trouver	88
10	rugby	52	voyage	101	spaghetti	100	cuir	79
11	brochette	51	nœud	100	voyage	91	dîner	78
12	luxe	51	spaghetti	98	frais	89	déguster	67
13	bille	51	enfant	97	long	87	voyage	62
14	poulet	50	grand	96	clé	86	perdre	56
15	voiture	50	partir	85	dos	83	sol	54
16	escalier	50	ranger	80	main	82	scintiller	54
17	marbre	49	invité	79	attraction	78	canard	53
18	argent	49	cravate	77	parfaire	74	dame	53
19	attraction	49	glisser	69	grand	71	dos	53
20	stylo	49	cacher	68	oublier	64	poser	52
21	aimer	47	main	67	soie	64	main	52
22	faire	47	préférer	67	poser	62	nuit	51
23	dos	47	hier	66	déguster	62	papillon	51
24	commission	46	match	64	canard	61	rugby	50
25	carbonara	41	frais	64	sol	60	volley	50
26	fêter	40	oublier	62	foie	59	poulet	50
27	voyage	40	dos	62	falloir	59	clef	50
28	nuit	40	lumière	59	cuir	59	voiture	50
29	jeu	39	adorer	58	stylo	57	entrecôte	50
30	noeud	37	poser	57	envier	56	frite	50
31	perdre	36	voiture	57	escalier	54	escalier	50
32	acheter	35	accrocher	57	remplir	54	marbre	50
33	volley	35	long	56	sport	54	foie	50
34	canard	34	croustillant	55	marbre	53	jeu	50
35	dame	34	porter	54	parent	52	argent	50
36	gras	33	bien	52	nuit	52	attraction	50
37	sol	32	stylo	52	brochette	51	huître	50
38	papillon	32	escalier	51	voiture	51	bille	50
39	mettre	31	marbre	51	préférer	51	commission	50
40	foie	31	foie	51	dame	51	luxe	50
41	adorer	30	canard	51	luxe	51	sport	50
42	aller	30	nuit	51	sucre	51	carbonara	50
43	bolognais	29	rugby	50	rugby	50	stylo	50
44	trou	29	gonfler	50	volley	50	sucre	50
45	cuir	27	volley	50	poulet	50	poudre	50
46	tomate	27	poulet	50	entrecôte	50	tomate	50
47	sport	26	entrecôte	50	frite	50	trou	50
48	souris	26	jeu	50	gras	50	brochette	49
49	bolognaise	25	argent	50	jeu	50	soir	49
50	plume	25	papillon	50	argent	50	cravate	49

Tableau 15 : Table de fréquence lexicale des 50 mots pleins lemmatisés les plus fréquents dans les phrases sollicitées et artificielles

Annexe 2 : Emploi de chaque séquence collocative nominale dans les phrases sollicitées et artificielles

SÉQUENCE	Humains	Ministral-3 (14b)	Gemma3 (27b)	GPT-oss (120b)
ballon de rugby	70,00 %	52,00 %	100,00 %	98,00 %
ballon de volley	35,29 %	28,00 %	36,00 %	76,00 %
brochette de poulet	74,00 %	100,00 %	100,00 %	100,00 %
clé de sol / clef de sol	52,78 %	2,04 %	18,00 %	0,00 %
clé/clef de voiture	46,00 %	0,00 %	64,00 %	0,00 %
clé/clef de la voiture	22,00 %	18,00 %	32,00 %	44,00 %
clé/clef de ma/ta/…/leur voiture	20,00 %	16,00 %	2,00 %	42,00 %
entrecôte avec des frites	38,10 %	0,00 %	26,00 %	0,00 %
entrecôte frites	4,76 %	0,00 %	0,00 %	0,00 %
escalier de marbre	14,29 %	4,00 %	100,00 %	12,00 %
escalier en marbre	34,69 %	90,00 %	0,00 %	72,00 %
foie gras de canard	36,67 %	38,00 %	96,00 %	0,00 %
foie de canard	0,00 %	0,00 %	4,00 %	44,00 %
jeu de dames	53,12 %	16,33 %	0,00 %	34,00 %
jeu des dames	3,12 %	0,00 %	0,00 %	0,00 %
noces d’argent	73,47 %	0,00 %	24,00 %	0,00 %
nuit de noces	79,49 %	0,00 %	6,00 %	2,00 %
nuit des noces	0,00 %	0,00 %	0,00 %	0,00 %
nuit de mes/tes/…/leurs noces	5,13 %	38,78 %	0,00 %	0,00 %
noces d’or	96,15 %	0,00 %	72,00 %	4,00 %
nœud de cravate	42,86 %	10,00 %	2,00 %	2,00 %
nœud de la cravate	0,00 %	0,00 %	0,00 %	8,00 %
nœud de ma/ta/…/leur cravate	19,05 %	86,00 %	22,00 %	54,00 %
nœud papillon	81,25 %	2,00 %	40,00 %	2,00 %
nœud de papillon	3,12 %	0,00 %	0,00 %	0,00 %
parc d’attractions	63,04 %	86,00 %	56,00 %	16,00 %
parc à huîtres	52,38 %	0,00 %	0,00 %	0,00 %
sac de billes	74,00 %	0,00 %	8,00 %	0,00 %
sac à billes	0,00 %	0,00 %	0,00 %	0,00 %
sac à commissions	8,16 %	0,00 %	0,00 %	0,00 %
sac de commissions	4,08 %	6,00 %	0,00 %	0,00 %
sac en cuir	60,87 %	98,00 %	98,00 %	100,00 %
sac de cuir	4,35 %	0,00 %	0,00 %	0,00 %
sac à dos	68,75 %	100,00 %	100,00 %	20,00 %
sac de luxe	66,00 %	4,08 %	10,00 %	30,00 %
sac à main	76,00 %	52,00 %	100,00 %	14,00 %
sac de sport	61,90 %	32,65 %	66,00 %	96,00 %
sac de voyage	31,91 %	12,00 %	4,00 %	16,00 %
spaghettis bolognaise	33,33 %	10,00 %	0,00 %	14,00 %
spaghettis à la bolognaise	31,25 %	76,00 %	92,00 %	72,00 %
spaghettis carbonara	29,79 %	18,00 %	0,00 %	6,00 %
spaghettis à la carbonara	40,43 %	72,00 %	20,00 %	82,00 %
stylo plume	29,79 %	0,00 %	16,00 %	2,00 %
stylo à plume	17,02 %	22,00 %	42,00 %	20,00 %
sucre en poudre	89,47 %	94,00 %	96,00 %	66,00 %
sucre poudre	5,26 %	0,00 %	0,00 %	0,00 %
tomate mozzarella	15,38 %	0,00 %	0,00 %	0,00 %
tomate à la mozzarella	3,85 %	0,00 %	0,00 %	0,00 %
trou de souris	53,85 %	0,00 %	0,00 %	0,00 %

Tableau 16 : Taux d’emploi de chaque séquence collocative nominale par chacun des participants

Sont représentés en vert tous les taux ≥ 50% ; en rouge < 50%