Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Embeddings, topic models, LLM : un air de famille

Ludovic Tanguy Cécile Fabre Nabil Hathout Lydia-Mai Ho-Dac
CLLE, Université de Toulouse, Toulouse, France
{prenom.nom}@univ-tlse2.fr

RÉSUMÉ

Cet article présente une étude portant sur les termes exprimant les relations familiales (*frère*, *tante*, etc.) à travers trois méthodes : les plongements de mots, le topic modeling et les modèles de langue. Les deux premières représentations sont construites sur la version française de Wikipédia, la troisième est obtenue par une interrogation directe de ChatGPT. L'objectif est de comparer les représentations de ces termes par les trois méthodes, et ce de deux façons : en les confrontant à une définition structurelle des relations familiales (en termes de traits comme le genre, l'ascendance, etc.) et en comparant les thématiques associées à chaque terme. Ces méthodes permettent d'identifier différents modes de structuration du vocabulaire de la famille, tout en montrant qu'un recours au corpus et à des analyses contrôlées reste indispensable pour aboutir à des résultats fiables.

ABSTRACT

Word embeddings, topic models, LLMs: a family affair

This article presents a study on terms denoting family relationships (*brother*, *aunt*, etc.) in French using three approaches: word embeddings, topic modeling, and pre-trained language models. The first two types of representations are built from the French version of Wikipedia, while the third is derived through direct interaction with ChatGPT. The aim is to compare how these three methods represent such terms, in two main ways: by evaluating them against a structural definition of family relations (in terms of features such as gender, lineage, etc.), and by comparing the topics associated with each term. These methods reveal different modes of structuring family-related vocabulary, while also underscoring the continued necessity of corpus-based and controlled analyses to obtain reliable results.

MOTS-CLÉS : Plongements de mots, topic modeling, LLM, lexique de la famille.

KEYWORDS: Word embeddings, topic modeling, LLMs, family lexicon.

1 Introduction

Les termes de parenté (comme *frère*, *mère*, *tante*, etc.) constituent un champ lexical très fréquent dans les textes, qui permet d'accéder à des informations riches sur les positions qu'occupent les individus en termes de rôles et de normes sociales. Comme d'autres champs lexicaux —noms de pays, de capitales ou de couleurs— ils forment des clusters distributionnels denses (Mikolov *et al.*, 2013), comportant un ensemble relativement réduit de membres rassemblés par leur appartenance à une classe ontologique bien délimitée. Le champ lexical de la famille se distingue néanmoins des autres par son caractère fortement structuré, puisqu'il s'organise selon des relations d'ascendance, de descendance, de germanité (descendants de mêmes parents), d'alliance, décrits de longue date

Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux, pages 295–312.

par les sciences sociales (Barry, 2018). D'autres niveaux de structuration existent certainement pour les décrire, comme ceux induits par les rôles socialement attribués aux différents membres de la famille mais ils sont moins évidents et moins formalisés. La structuration par les relations de parenté découle directement du sens des termes familiaux comme le fait que *mère* désigne 'un géniteur de genre féminin'. Les autres structurations sont en revanche déterminées par l'usage, reflet des rôles et des attributions des individus qui composent les familles. L'un des objectifs de notre étude est de mettre en évidence quelques-unes des propriétés qui organisent les familles mais ne peuvent pas être déduites des seules relations de parenté.

Nous proposons d'identifier ces relations et ces qualités au moyen de trois méthodes dont nous cherchons à comparer la pertinence vis-à de cet objectif. La première est fondée sur la similarité paradigmatique, et consiste à décrire les thématiques associées aux termes familiaux à partir d'une annotation manuelle de leurs voisins distributionnels calculés via des embeddings statiques. La deuxième exploite, dans une approche syntagmatique, les propriétés sémantiques des cooccurrents des termes familiaux, en ayant recours à une méthode de *topic modeling* (LDA) pour thématiser les énoncés dans lesquels apparaissent les termes de la familles. La troisième méthode consiste à demander directement à un LLM de proposer une caractérisation thématique des termes familiaux. La comparaison des résultats obtenus par ces 3 méthodes montre que les deux premières approches sont assez fortement corrélées, avec des résultats proches vis-à-vis des traits distinctifs de la parenté et des thématiques sémantiques qui émergent. Les représentations fournies par le LLM s'écartent fortement des précédentes, et montrent une grande versatilité. Ce travail a été mené sur le français à partir du corpus des articles Wikipedia, ce qui facilitera des études comparatives multilingues.

Après la présentation de travaux connexes, nous décrivons en section 3 le matériau sur lequel se base notre étude : le corpus, le lexique et les méthodes utilisées pour représenter les termes exprimant les relations familiales. La suite de l'article est consacrée à la comparaison de ces représentations et de la manière dont elles organisent les termes : en prenant comme référence l'organisation conceptuelle de la famille (section 4) et en identifiant les thématiques associées à chaque terme (section 5).

2 Travaux connexes

Cette étude s'inscrit dans la lignée de travaux comme ceux de Kozlowski *et al.* (2019) et Erk & Apidianaki (2024) qui s'intéressent à la capacité des modèles computationnels, et en particulier des embeddings de mots, à capter des informations sémantiques riches révélatrices de dimensions sociales et culturelles. Le vocabulaire de la famille est un bon candidat car il est porteur de relations stéréotypiques entre les genres et les générations. Peu de travaux de linguistique de corpus ont porté sur ce champ lexical, hormis, à notre connaissance, Čermáková & Mahlberg (2021), qui s'intéressent aux noms de famille selon l'angle du genre dans des textes littéraires pour enfants. Elles montrent comment l'étude comparée des noms *father* et *mother* est révélatrice d'usages linguistiques et de représentations sociales très contrastées.

Dans la présente étude, nous mobilisons des méthodes qui permettent de sonder à plus grande échelle ces représentations. La première est similaire à celle de Wauquier *et al.* (2020) qui montrent comment les outils d'analyse distributionnelle permettent d'étudier de façon systématique l'organisation sémantique de classes de mots, en l'occurrence de noms d'action. Huyghe & Wauquier (2021) proposent une étude comparable sur des noms d'agent. Notre deuxième méthode s'inspire de celle de Wanna *et al.* (2024) qui ont fourni à des LLM les sorties d'une analyse de type *topic modeling* pour générer

des étiquettes décrivant les topics, et qui les ont comparées manuellement avec celles produites par des experts, avec un score de similarité satisfaisant. Gillings & Jaworska (2025) ont pour leur part comparé différentes techniques d'identification de thématiques dans de petits corpus, et notamment l'étiquetage des topics produits par une analyse LDA, par des LLM et par des humains utilisant différentes méthodes d'exploration, mais sans conclure sur l'existence d'un accord ou d'un désaccord. Notre troisième méthode est comparable à celle de Uchida (2024) qui a testé la capacité des LLM génératifs à produire des listes de mots fréquents et des collocations courantes pour l'anglais, et comparé ces résultats à un corpus de référence (COCA) avec un accord probant pour les phénomènes les plus fréquents.

3 Données et modèles

3.1 Corpus et prétraitements

Notre étude se base sur un corpus construit à partir du code HTML de la version française de Wikipédia (dump du 1er décembre 2018) qui contient 1.4 millions d'articles et 720 millions de mots. Nous avons préféré cette version un peu ancienne au vu de la qualité du texte (basé sur le code HTML rendu et non sur le wikicode qui pose un ensemble de problèmes). Nous avons créé une version lemmatisée et neutralisée en genre grammatical. Nous avons utilisé l'analyseur Stanza (Qi *et al.*, 2020) version 1.3 avec le modèle UD_French-GSD pour la segmentation, l'étiquetage morphosyntaxique et une première lemmatisation. La lemmatisation produite par Stanza n'étant pas toujours de qualité suffisante (contrairement à l'analyse morphosyntaxique), nous l'avons corrigée en utilisant le lexique flexionnel Morphalou 3 (ATILF, 2023). La lemmatisation est en effet une étape critique dans notre approche pour identifier les mots-cibles et leurs cooccurrents. Pour les classes ouvertes (noms, verbes, adjectifs et adverbes), nous avons identifié dans Morphalou (ou prédit à partir de la forme) le lemme correspondant à la forme de surface, à la catégorie et aux traits flexionnels proposés par Stanza. Pour les classes fermées et en cas d'absence d'une forme de classe ouverte compatible dans le lexique nous avons gardé la lemmatisation proposée par Stanza. Afin de limiter les biais liés au genre, toutes les marques du genre grammatical ont été supprimées. Le modèle GSD de Stanza lemmatise déjà les déterminants et les pronoms par la forme masculine. Pour la même raison, nous avons remplacé les noms propres, notamment les prénoms, par une chaîne générique "NAM" pour réduire les biais qu'ils induisent. Par exemple, la phrase en (1) est lemmatisée comme en (2).

(1) Austin, le neveu d'Edie emménage chez sa tante et entame rapidement une relation avec Julie, la fille de Susan.
(2) NAM, le neveu de NAM emménager chez son tante et entamer rapidement un relation avec NAM, le fille de NAM.

3.2 Choix des amorces

Nous avons sélectionné une liste de noms de parenté en partant d'un recensement disponible sur Wikipédia [1]. Cette liste initiale a été réduite à 25 noms (3) après élimination des termes dont la fréquence est inférieure à un seuil fixé à 1000 dans notre corpus. Ce seuil de fréquence permet

1. https://fr.wikipedia.org/wiki/Parenté#Dénominations

d'obtenir des représentations correctes et stables pour chaque mot en termes de plongements lexicaux, l'homogénéité de la gamme de fréquence permettant par ailleurs des comparaisons plus fiables entre les termes.

(3) beau-frère, beau-père, belle-fille, belle-mère, belle-sœur, cousin, cousine, épouse, époux, fille, fils, frère, gendre, grand-mère, grand-père, mari, mère, neveu, nièce, oncle, père, petit-fils, petite-fille, sœur, tante

Les termes écartés correspondent aux relations les plus éloignées (ex : *arrière-grand-mère*, *petite-nièce*, *demi-frère*, etc.) ou à des variantes moins usitées (*gendre* est bien plus fréquent que *beau-fils*, et *belle-fille* que *bru*). Nous avons également filtré la liste de départ afin de garder les sous-ensembles complets correspondant à des types de relations familiales comme la belle famille ou les ascendants et descendants de même génération. Nous avons choisi d'écarter le terme *femme* (en tant que correspondant féminin de *mari*) dont le sens majoritaire dans le corpus est celui de 'individu de sexe féminin'. Sa prise en compte aurait faussé nos résultats. Nous avons en revanche conservé *fille* dont l'acception 'femme jeune' n'est pas dominante au vu des différents modèles construits.

3.3 Représentations issues des deux méthodes sur corpus

Plongements de mots. Le premier type de représentation des termes que nous avons étudié est constitué par des plongements de mots statiques, autrement dit des représentations vectorielles calculées au niveau du type. Ce choix d'une technologie plus ancienne par rapports aux plongements contextuels produits au niveau des occurrences (notamment par des architectures transformers de type BERT) se justifie de plusieurs manières. Tout d'abord, nous cherchons à obtenir une représentation synthétique de l'emploi de chaque terme dans le corpus, ce qui est la nature des plongements statiques, même si cela conduit à écraser les variations sémantiques et la polysémie. Les plongements contextuels nécessitent dans notre cas une agrégation à posteriori (typiquement une moyenne) qui finit par les appauvrir à grande échelle. Les plongements statiques ont d'ailleurs montré leur efficacité dans plusieurs tâches comme le repérage du changement sémantique (Schlechtweg, 2023). Enfin, les plongements contextuels nécessitent des choix critiques sur les couches cachées utilisées, ce qui a un impact crucial et peut facilement faire baisser leur efficacité sur des tâches de sémantique lexicale (Miletic & Schulte im Walde, 2023).

Nous avons donc construit des plongements de mots statiques des 25 amorces (3) en utilisant l'algorithme Skip Gram with Negative Sampling (SGNS), popularisé par Word2vec de Mikolov *et al.* (2013), implémenté dans la bibliothèque Gensim [2] (Řehůřek & Sojka, 2010). L'algorithme SGNS, comme toutes les méthodes d'apprentissage basées sur des réseaux de neurones, fait intervenir un ensemble de processus aléatoires, ce qui entraîne un non-déterminisme et de ce fait une certaine instabilité des modèles vectoriels. Afin de limiter l'impact de ces aléas, il est d'usage de générer plusieurs modèles avec les mêmes données et paramètres, et de les utiliser conjointement (Pierrejean & Tanguy, 2018). Nous avons donc entraîné 5 modèles et construit une représentation vectorielle unique en concaténant les 5 vecteurs de chaque mot : l'espace vectoriel est donc de dimension 1500.

Topic models. Nous avons également construit des *topic models* afin d'obtenir une représentation des contextes d'apparition de nos amorces. Notre objectif est d'avoir une caractérisation synthétique pour chaque amorce qui agrège les principales thématiques et classes lexicales. Pour cela, nous

2. Les paramètres utilisés sont : 300 dimensions, fréquence minimale de 100 occurrences, fenêtre de 5 mots, taux d'exemples négatifs 5, sous-échantillonnage 10^{-3}, 5 epochs.

avons choisi de considérer comme unités les phrases et non les documents afin que la variété des emplois des termes de la famille ne soit pas noyée dans l'homogénéité des articles de Wikipédia où ils apparaissent (biographies, résumés d'œuvres de fiction, etc.). Pour compenser les effets de fréquence, nous avons construit un sous-corpus semi-aléatoire composé de 34 150 phrases (telles que délimitées par Stanza, *cf supra*), soit 1366 pour chaque amorce, (1366 est le nombre de phrases dans lesquelles apparaît *belle-sœur*, l'amorce la moins fréquente). L'échantillonnage est réalisé avec remise. Le taux de doublons reste faible avec moins de 2% de phrases répétées parce qu'elles contiennent plusieurs amorces.

Nous avons choisi l'algorithme canonique LDA (Latent Dirichlet Allocation, Blei (2012)) tel qu'implémenté dans Mallet (McCallum, 2002) qui utilise une méthode d'échantillonnage (Gibbs) dont les résultats sont qualitativement plus satisfaisants que ceux de la version de Gensim. Nous avons appliqué la même segmentation que pour les plongements de mots, mais ajouté une stop-liste basée sur la liste générique proposée par Mallet pour le français, à laquelle nous avons ajouté le joker "NAM" et tous les termes de la famille (soit la totalité de notre liste de départ). Nous avons généré arbitrairement des modèles avec 50 dimensions (i.e. *topics*). Pour chaque amorce, nous avons construit sur cette base un vecteur qui correspond à la somme des poids de chaque topic sur l'ensemble des phrases qui la contiennent. Enfin, nous avons, sur le même principe que pour les plongements de mots, lancé le même processus 5 fois pour neutraliser les facteurs aléatoires et concaténé les vecteurs (donc de dimension 250). Nous proposons une analyse qualitative des thématiques identifiées en Section 5.2.

3.4 Représentations proposées par un LLM

Nous avons soumis une série de requêtes à ChatGPT (version 4o) afin d'obtenir une matrice dont les lignes sont constituées des 25 amorces et dont les colonnes comportent les catégories sémantiques identifiées par le LLM comme celles qui caractérisent le mieux les amorces. Le score attendu est celui correspondant au degré d'association estimé entre chaque mot et chaque catégorie (entre 0 et 10). Constatant que le LLM avait tendance à proposer principalement des thématiques définitoires de la parenté (alliance, ascendance, genre), nous avons soumis des prompts présentant des variations de formulation afin de faire émerger des thématiques plus diversifiées. Par exemple, au lieu de parler de thématiques qui *distinguent* les mots les uns des autres, nous avons parlé de thématiques *évoquées* par les amorces. Nous avons également fait varier le nombre de catégories attendues, qui restent comprises entre 12 et 20. Dans tous les cas, la référence aux usages de ces mots dans la Wikipédia est explicitement indiquée dans le prompt. Quatre modèles ont été retenus (dont les matrices sont identifiées par la suite ChaGPT-1 à ChatGPT-4). Le détail des thématiques générées et un exemple de prompt sont donnés en annexe A.3.

Bien que conscients de l'opacité de ce LLM propriétaire, nous avons décidé de l'utiliser pour son statut emblématique. Les expériences menées avec des modèles ouverts de petite taille (mistral, llama3 etc.) nous semblent produire des résultats qualitativement très proches, sans fournir des informations nécessairement plus précises sur les données d'entraînement.

4 Comparaison des différentes représentations

Dans cette section nous comparons les différentes représentations obtenues des 25 amorces, dans un première temps entre elles, puis en les confrontant à la structure inhérente aux concepts de la famille.

4.1 Distance entre les amorces

Pour comparer les différentes représentations des amorces, nous avons calculé pour chaque modèle une matrice de distance entre les amorces en utilisant la similarité cosinus ($dist(x, y) = 1 - sim_{cos}(x, y)$). Nous avons appliqué cette mesure séparément sur les ensembles de vecteurs décrits précédemment (embeddings, poids des topics et coefficients proposés par le LLM). Puis, sur les N=300 paires d'amorces à considérer, nous avons mesuré la corrélation entre ces distances. La table 1 donne les valeurs des coefficients de corrélation de Spearman (ρ) pour chaque paire de modèles. On y voit que si les distances entre les représentations basées sur les plongements de mots et celles de la LDA sont relativement similaires ($\rho = 0.61$), celles produites par ChatGPT montrent une importante variation. Parmi les 4 variantes retenues, seules les deux premières apparaissent cohérentes entre elles et partiellement corrélées avec les plongements de mots.

Modèle	SGNS	LDA	ChatGPT-1	ChatGPT-2	ChatGPT-3
LDA	0.61				
ChatGPT-1	0.35	0.01			
ChatGPT-2	0.21	0.00	0.54		
ChatGPT-3	-0.03	0.02	-0.01	-0.09	
ChatGPT-4	-0.01	-0.05	-0.04	-0.12	0.13

TABLE 1 – Matrice de corrélation (Spearman) des distances entre amorces selon chaque modèle

4.2 Lien avec les traits structurels de la famille

Dans un deuxième temps, nous avons voulus savoir si la représentation des amorces dans chaque modèle suivait la structure interne qui régit les relations familiales. Autrement dit, dans quelle mesure la distance qui sépare les vecteurs de deux amorces est-elle proportionnelle à la différence entre les concepts ? Pour estimer cette dernière, nous nous sommes basés sur une décomposition en traits distinctifs des différents termes, telle que la proposent traditionnellement les anthropologues (Segalen & Martial, 2019). Nous avons retenu une décomposition en 5 traits qui permet de décrire de façon unique chaque relation familiale en considérant comment un membre de la famille est rattaché à un point de vue central (égo) : le Genre, l'Ascendance (1 pour *père*, 2 pour *grand-mère*), la Descendance (1 pour *fille*, 2 pour *petit-fils*), l'Alliance (*époux*, *belle-sœur*) et la Germanité (*frère*, *cousine*). Par exemple *père* se décompose en Masculin, Ascendance (degré 1), *nièce* se décompose en Féminin, Descendance (degré 1), Germanité. Il existe peu d'ambiguïtés dans notre liste de termes. Si *beau-frère* peut désigner soit le conjoint du frère ou de la sœur, soit le frère du conjoint, cela ne change pas sa décomposition en traits (Masculin, Germanité, Alliance). Par contre *tante* peut désigner l'épouse de l'oncle (du frère du parent) ou la sœur du parent. Dans ce cas nous avons assigné un score intermédiaire (0.5) pour le trait Alliance.

Il est possible de décrire chaque paire d'amorces en termes de différence de traits. Par exemple, entre *frère* et *soeur* seul le trait Genre varie. Entre *frère* et *grand-mère* il existe au contraire une différence de Genre, d'Ascendance (2 degrés) et de Germanité (trait présent pour *frère* et absent pour *grand-mère*).

Nous avons donc comparé, pour chaque modèle, cette distinction en traits avec la distance entre les vecteurs des deux termes. Nous avons pour cela construit pour chaque représentation vectorielle, sur la base des 300 paires d'amorces, un modèle statistique de régression linéaire multiple en prenant

comme variable dépendante la distance vectorielle et comme variables indépendantes (ou prédicteurs) les différences de traits entre les deux termes de la paire.

Deux types d'information peuvent être obtenus à partir de ces différents modèles linéaires. Le premier est le R^2 ajusté, autrement dit la part de variance de la distance dans l'espace vectoriel qui peut être expliquée par la différence des traits structurels. Le second est l'importance relative de chacun de ces traits. Nous l'avons mesurée en utilisant une méthode par ablation, c'est-à-dire en calculant la part du R^2 ajusté perdue par le modèle linéaire lorsqu'on supprime ce trait. Par exemple, si l'on supprime le trait Ascendance, il n'y a plus de différence structurelle entre *grand-mère* et *mère* pour justifier la distance entre les deux termes.

Modèle	R^2	Genre	Ascendance	Descendance	Germanité	Alliance
SGNS	0.36	-0.50	-0.34	-0.18	-0.06	-0.10
LDA	0.34	-0.28	-0.07	-0.75	0.00	0.00
ChatGPT-1	0.39	0.00	0.00	0.00	-0.49	-0.53
ChatGPT-2	0.77	0.00	0.00	0.00	-0.27	-0.75
ChatGPT-3	0.00	NA	NA	NA	NA	NA
ChatGPT-4	0.00	NA	NA	NA	NA	NA

TABLE 2 – Analyse des régressions linéaires pour prédire la distance vectorielle à partir des traits distinctifs : R^2 ajusté total du modèle et perte relative par ablation de chaque trait

La table 2 donne le détail de ces informations. On peut y voir que les distances calculées à partir du corpus (SGNS et LDA) ne sont expliquées qu'à hauteur de 35% par les différences de traits entre les amorces. En revanche, les traits les plus distinctifs varient entre les deux méthodes : pour SGNS ce sont les différences de genre qui sont les plus importantes, puis celles sur le trait Ascendance, et plus faiblement Descendance. Pour LDA on retrouve ces mêmes 3 traits mais avec des poids différents, celui lié à Descendance étant le plus important, avec une baisse de 75% du R^2 si on supprime ce trait.

Les représentations proposées par ChatGPT montrent d'importantes différences et variations. Les deux premières (ChatGPT-1 et ChatGPT-2) indiquent que les traits descriptifs proposés semblent en cohérence avec la décomposition structurelle des relations familiales, notamment pour la deuxième qui atteint un R^2 de 77%. De fait, certaines des thématiques descriptives proposées reprennent presque littéralement les traits structurels. On notera toutefois que ce sont les traits Germanité et Alliance qui sont mobilisés, et non comme pour les deux modèles précédents les traits correspondant au Genre et à l'Ascendance/Descendance. De fait, il apparaît que dans les deux premières propositions générées par ChatGPT, les différences de genre sont inexistantes : les représentations sont strictement identiques pour les paires genrées (*frère* - *sœur*, *cousin* - *cousine* etc.). Une hypothèse pour expliquer ce fait est que les mécanismes d'alignement de ces modèles (*instruction tuning*) ont spécifiquement cherché à réduire, voire à annuler les biais de genre qui ont très vite été identifiés et critiqués dans ces modèles (Bolukbasi *et al.*, 2016). Il semblerait que l'alignement des modèles conversationnels grand public ait atteint une phase où les différences de genre sont systématiquement annulées dans les sorties, quand bien même elles existent dans les données d'apprentissage (Chen *et al.*, 2025).

Les deux dernières représentations proposées par ChatGPT (ChatGPT-3 et ChatGPT-4), alors qu'elles présentent des thématiques qui semblent du même ordre que les deux premières (voir annexe A.3), mènent à des distances entre les amorces qui n'ont absolument aucun lien avec la décomposition en traits structurels (R^2 nuls, rendant inapplicable l'ablation). De fait, un examen manuel des poids des thématiques montre un ensemble important d'incohérences, qui reviennent à une génération aléatoire

des valeurs dans la matrice. À des fins d'illustration, la figure 1 montre une projection t-SNE de trois des modèles vectoriels (basée sur la même distance que précédemment), avec une mise en relief des traits structurels principaux, qui montrent bien que l'organisation interne des termes varie de façon importante d'une représentation à l'autre [3].

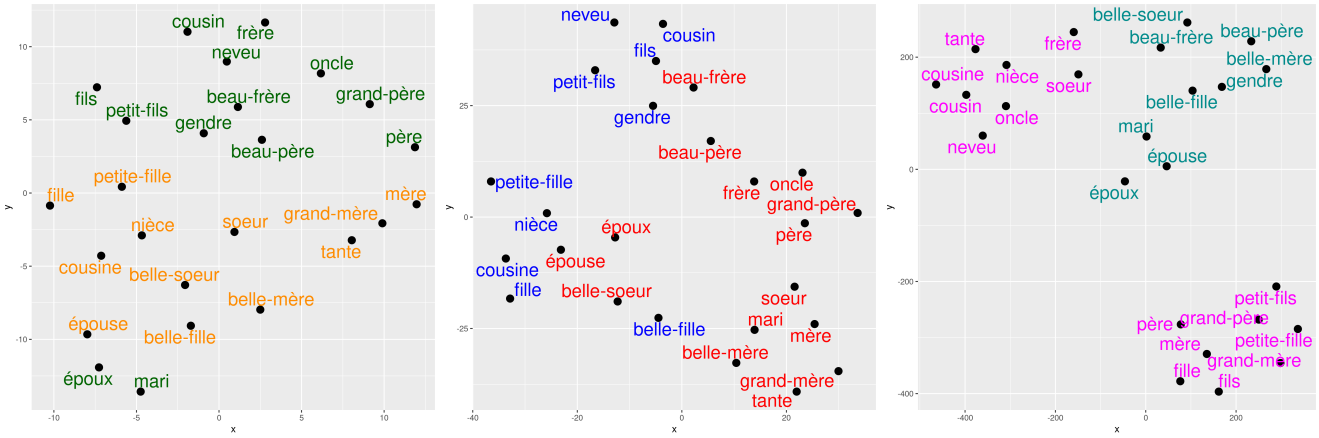

FIGURE 1 – Projections t-SNE : SGNS avec marquage du trait Genre (à gauche), LDA avec marquage du trait Descendance (au centre) et ChatGPT-1 avec marquage du trait Alliance (à droite)

5 Comparaison des thématiques associées aux amorces

Pour poursuivre l'analyse et explorer la part de variance qui n'est pas explicable par les traits structurels, il est nécessaire d'aborder de façon plus qualitative les représentations en caractérisant ce qui distingue les différents termes dans chaque modèle. Étant donné la différence entre les techniques mobilisées, nous avons abordé chacune de façon distincte.

5.1 Thématiques des embeddings : annotation manuelle des voisins distributionnels

Afin d'identifier les thématiques associées à chaque amorce d'après les plongements de mots calculés par la méthode SGNS, nous faisons l'hypothèse que les caractéristiques sémantiques des mots proches de chaque amorce sont une bonne approximation des traits sémantiques de l'amorce elle-même. Nous avons réalisé une annotation des plus proches voisins distributionnels de chaque amorce, toujours en nous basant sur le cosinus entre les vecteurs. Cette annotation a été effectuée manuellement, suivant la procédure que nous décrivons ici.

Avant tout, il est important de noter à ce stade la grande cohésion du lexique de la famille. Quelle que soit l'amorce, les autres termes de la famille se retrouvent systématiquement parmi les plus proches voisins : 80% au rang 10, 60% au rang 50. Afin de pouvoir les considérer plus globalement, nous avons donc extrait l'ensemble des 200 plus proches voisins de chaque amorce, soit 590 mots distincts. Les auteurs de l'article ont ensuite examiné chacun un échantillon de ces voisins, avec comme objectif de proposer des thématiques permettant de les regrouper. Les différentes propositions

3. La distance entre les mots d'une paire genrée dans le modèle ChatGPT-1 est en réalité nulle puisque les vecteurs sont identiques. Ce sont les mécanismes stochastiques de la projection t-SNE qui induisent un léger éloignement des points.

ont été rassemblées puis discutées pour arriver à un consensus sur 21 thématiques non exclusives (après avoir écarté les termes désignant des relations familiales, dont les amorces et les autres termes de la liste initiale). Le processus d'annotation final a été mis en place en distribuant les voisins parmi les différents juges, après avoir vérifié que l'accord inter-annotateur était élevé : sur un échantillon de 133 voisins annotés par 2 à 4 juges, nous avons mesuré un coefficient α de Krippendorff [4] global de 0.82.

La liste des thématiques et des exemples de mots associés est présentée en annexe A.1. Elles couvrent des champs sémantiques très variés, avec, parmi les thématiques les plus représentées : les termes désignant un titre ou une fonction (*chevalier*, *nonne*), un notable ou un noble (*comtesse*, *tsar*), un lien ou une relation (*ami*, *collègue*), un nom de métier (*actrice*, *rentier*), un sentiment (*jalousie*, *chéri*).

5.2 Thématiques du LDA : étiquetage des topics par un LLM

Pour décrire les thématiques identifiées par le topic model, nous avons décidé de recourir à un étiquetage automatique par un LLM. Cette pratique est devenue courante et donne des résultats satisfaisants (Wanna *et al.*, 2024). L'annotation s'est faite en considérant, pour chaque topic (50 pour chacun des 5 modèles), la liste des 50 mots les plus fortement associés par ordre décroissant de probabilité.

Nous avons soumis chacun des 5 ensembles de 50 topics à ChatGPT (version 4o) avec pour consigne d'assigner une thématique à chaque topic, et en encourageant l'utilisation des mêmes étiquettes pour les topics semblables, ainsi que le recours à une étiquette BRUIT pour les cas les plus hétérogènes. Nous avons réalisé 2 tests par modèle, générant 500 étiquettes, soit 98 étiquettes uniques. Afin de limiter leur dispersion, nous avons réduit manuellement les variations formelles (alternance singulier pluriel, variation de séparateurs dans le cas de thématiques avec plusieurs mots), puis examiné l'ensemble des étiquettes qui n'avaient été générées que par un modèle. Après consultation du topic correspondant, nous avons opté pour leur maintien, leur remplacement par l'étiquette BRUIT, ou leur assimilation à une étiquette sémantiquement équivalente plus fréquente. Nous avons finalement obtenu 67 étiquettes, que nous avons soumises au LLM pour générer des catégories plus larges les regroupant. Après plusieurs tests consistant à estimer manuellement le nombre de catégories optimal au vu de la variété des topics, nous avons retenu l'option à 17 catégories (et une catégorie résiduelle DIV), que nous avons utilisées pour étiqueter le premier des cinq modèles.

Des exemples de prompts utilisés ainsi que les étiquettes retenues pour ce modèle sont indiqués en annexe A.2 avec quelques-uns des mots avec les poids les plus forts pour les topics correspondants.

5.3 Comparaison des thématiques entre les modèles

Pour chaque modèle, nous disposons d'une matrice contenant, pour chaque amorce et pour chaque thématique, un score correspondant à la force (ou pertinence) de l'association de cette thématique à cette amorce. Pour les plongements de mots, il s'agit du nombre de voisins distributionnels relevant de la thématique parmi les 200 premiers ; pour le topic model il s'agit du poids cumulé des topics correspondants sur l'ensemble des phrases contenant l'amorce (en nombre égal pour chaque

4. Le choix de cette mesure de l'accord est justifié par le fait que la répartition des items entre les annotateurs était irrégulière. Sur les sous-ensembles de voisins annotés par les mêmes annotateurs les kappa de Cohen et de Fleiss sont également proches de 0.8.

dans l'échantillon) ; pour les modèles produits directement par ChatGPT il s'agit du score fourni directement par le LLM.

Les différents jeux de thématiques ne sont donc pas directement comparables, car les catégories identifiées tout comme les étiquettes qui les décrivent proviennent de processus indépendants et appliqués à des données différentes.

Afin de permettre un premier repérage des points de convergence ou de divergence, nous avons calculé pour chaque matrice le coefficient de corrélation de Spearman entre les traits structurels décrivant les amorces (Genre, Alliance, Descendance, Ascendance, et Germanité) et chacune des thématiques. Nous avons ensuite isolé, pour chacun de ces traits, les thématiques qui lui sont significativement corrélées (au seuil de 0.05, ddl=23 pour N=25 amorces). La table 3 montre les thématiques ainsi identifiées pour les plongements de mots et le topic modeling.

Trait	SGNS	LDA	ChatGPT (tous)
Ascendance	âge, enfance, foyer, lien/relation, métier/occupation, santé	Arts & Culture, Quotidien & Environnement, Relations & Sentiments	mythologie, autorité, histoire et société
Descendance	armée, notable/noble, titre/fonction	Histoire & Héritage, Personnalités & Figures Historiques, Pouvoir & Société	transmission, autorité, sociologie
Germanité	crime/délit, religion		cinéma
Alliance	crime/délit, sexe	Conflits & Drames	
Genre féminin	art, caractéristique, enfance, foyer, mort, procréation, sentiment/amour	Célébrations & Rites, Famille & Origines, Naissance & Mort, Quotidien & Environnement, Relations & Sentiments	héritage
Genre masculin	crime/délit, héritage, notable/noble, religion, titre/fonction	Conflits & Drames, Économie & Argent, Éducation & Transmission, Histoire & Héritage, Pouvoir & Société, Sports & Loisirs	

TABLE 3 – Thématiques significativement corrélées aux traits structurels, par modèle

On voit ici confirmé le fait que les représentations produites directement par ChatGPT sont éloignées de celles des deux autres modèles. Elles sont par ailleurs très limitées en nombre de thématiques. Nous nous concentrons donc sur les modèles SGNS et LDA, dont les thématiques présentent un recouvrement non négligeable. Autrement dit, une partie des associations se retrouvent dans les rapprochements paradigmatiques issus du modèle distributionnel comme dans ceux, syntagmatiques, qui proviennent du topic modeling. Pour le trait Ascendance, on retrouve les thématiques des relations et du quotidien (foyer, enfance, santé) ; pour le trait Descendance, les thématiques du pouvoir, et des personnalités publiques ; pour le trait Alliance, des thématiques à polarité négative (crime, conflit). Les différences de genre sont également marquées de façon identique : les termes féminins sont associés à des thèmes autour des sentiments, des étapes de la vie (naissance, enfance, mort) et du quotidien. A l'opposé, les termes masculins sont associés à des thématiques liées au pouvoir et à l'argent. Ces associations genrées sont celles que l'on sait exister massivement dans les corpus, et que les techniques de TAL font ressortir, y compris les LLM génératifs qui jusqu'à récemment tendaient à les amplifier (Stanczak & Augenstein, 2021). Le trait Germanité est par contre peu productif, mais rappelons qu'il recouvre un ensemble de termes disparates qui apparaissent dans des sphères assez séparées de l'espace familial (*frère*, *cousine*, *oncle*, *belle-sœur*...).

Concernant les différences entre les deux modèles, il faut noter que l'analyse distributionnelle fait surtout ressortir des formes nominales (professions, titres, et autres noms d'humains), alors que le topic model basé sur la cooccurrence génère au même titre des verbes, adjectifs et plus généralement des associations lexicales plus variées. On y retrouve ainsi, en observant les mots par lesquels les thématiques se manifestent, des ensembles souvent disjoints. Par exemple, le LDA va regrouper des termes liés à la fiction (littérature, mise en scène, cinéma) issus des contextes narratifs, alors que les plongements de mots vont faire ressortir des classes plus générales liées aux noms d'humains, substituables aux amorces dans ces mêmes contextes.

Nous avons mis en annexe B le premier plan factoriel résultant d'une analyse en composantes principales pour les modèles SGNS et LDA, ce qui permet d'avoir une vue d'ensemble des positions relatives de toutes les thématiques et amorces. Dans les deux cas, on retrouve bien les différences de genre (axes horizontaux, féminin à droite).

6 Conclusion et perspectives

Nous avons étudié le lexique de la famille en croisant trois façons d'obtenir des représentations sémantiques. Si les logiques et les mécanismes mobilisés par les trois méthodes sont incomparables, il reste possible d'identifier plusieurs points de convergence notamment entre les embeddings et le topic modeling. Le rôle du genre est prépondérant dans la structuration de ce lexique pour les méthodes basées sur corpus : les thématiques qui sont associées aux deux sous-ensembles se distinguent selon des dimensions socio-culturelles (santé, éducation, quotidien et sentiments pour les femmes ; argent et pouvoir pour les hommes). Les distinctions générationnelles sont elles aussi structurantes mais elles donnent des résultats moins nets dans la qualification. Ces résultats nous encouragent à approfondir l'étude de la complémentarité de ces deux approches pour mettre en évidence la structuration sémantique des classes de mots. Nous prolongerons ce travail en l'appliquant à l'anglais et à des corpus diachroniques pour examiner l'évolution des thématiques mises au jour.

L'approche directe par un LLM à qui on demande de se prononcer sur ce lexique confirme que si ces modèles sont bien, comme les deux autres, basés sur l'observation de régularités en corpus, celles-ci restent difficiles d'accès. C'est ce qu'ont montré quelques travaux récents [5] qui soulèvent la question de la nécessité de recourir à des corpus et à des analyses contrôlées pour identifier des tendances fiables, stables et interprétables. Mais il ne faut pas non plus considérer que les réponses d'un LLM à un prompt métalinguistique reflètent fidèlement les représentations construites par celui-ci (Hu & Levy, 2023).

En ce sens, une étude des embeddings contextuels produits par les modèles transformers est bien entendu nécessaire, mais elle soulève un ensemble de questions : comment neutraliser les contextes d'emploi comme nous l'avons fait ici, et, au-delà, comment contrôler l'influence respective du corpus de pré-entraînement et de projection (fine-tuning) ? Ce dernier point limite la possibilité d'étudier finement l'impact des sources textuelles, un facteur que l'on sait crucial et qui fait l'intérêt d'un champ lexical comme celui de la famille qui traverse les genres et les époques.

Remerciements

Les traitements ont été en partie effectués sur la plateforme OCCIDATA administrée par l'IRIT et soutenue par le CNRS et l'Université de Toulouse (https://occidata.irit.fr).

5. https://www.english-corpora.org/ai-llms/

Références

ATILF (2023). Morphalou. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

BARRY L. (2018). Logiques terminologiques. les taxinomies de parentéet leur relation aux systèmes d'alliance. *L'Homme. Revue française d'anthropologie*, **225**, 27–72.

BLEI D. M. (2012). Probabilistic topic models. *Communications of the ACM*, **55**(4), 77–84. DOI : 10.1145/2133806.2133826.

BOLUKBASI T., CHANG K.-W., ZOU J. Y., SALIGRAMA V. & KALAI A. T. (2016). Man is to computer programmer as woman is to homemaker ? debiasing word embeddings. *Advances in neural information processing systems*, **29**.

ČERMÁKOVÁ A. & MAHLBERG M. (2021). The representation of mothers and the gendered social structure of nineteenth-century children's literature. *English Text Construction*, **14**(2), 119–149.

CHEN E., ZHAN R.-J., LIN Y.-B. & CHEN H.-H. (2025). From structured prompts to open narratives: Measuring gender bias in llms through open-ended storytelling.

ERK K. & APIDIANAKI M. (2024). Adjusting interpretable dimensions in embedding space with human judgments. In K. DUH, H. GOMEZ & S. BETHARD, Éds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, p. 2675–2686, Mexico City, Mexico: Association for Computational Linguistics. DOI : 10.18653/v1/2024.naacl-long.146.

GILLINGS M. & JAWORSKA S. (2025). How humans and machines identify discourse topics: A methodological triangulation. *Applied Corpus Linguistics*, **5**(1), 100121. DOI : https://doi.org/10.1016/j.acorp.2025.100121.

HU J. & LEVY R. (2023). Prompting is not a substitute for probability measurements in large language models. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 5040–5060, Singapore: Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.306.

HUYGHE R. & WAUQUIER M. (2021). Distributional semantics insights on agentive suffix rivalry in French. *Word Structure*, **14**(3), 354–391.

KOZLOWSKI A. C., TADDY M. & EVANS J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, **84**(5), 905–949.

MCCALLUM A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. & DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, p. 3111–3119.

MILETIC F. & SCHULTE IM WALDE S. (2023). A systematic search for compound semantics in pretrained BERT architectures. In A. VLACHOS & I. AUGENSTEIN, Éds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 1499–1512, Dubrovnik, Croatia: Association for Computational Linguistics. DOI : 10.18653/v1/2023.eacl-main.110.

PIERREJEAN B. & TANGUY L. (2018). Predicting word embeddings variability. In *The seventh Joint Conference on Lexical and Computational Semantics*, p. 154–159.

QI P., ZHANG Y., ZHANG Y., BOLTON J. & MANNING C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

ŘEHŮŘEK R. & SOJKA P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, p. 45–50, Valletta, Malta: ELRA. http://is.muni.cz/publication/884893/en.

SCHLECHTWEG D. (2023). *Human and computational measurement of lexical semantic change*. Thèse de doctorat, Universität Stuttgart.

SEGALEN M. & MARTIAL A. (2019). La parenté des anthropologues. In M. SEGALEN & A. MARTIAL, Éds., *Sociologie de la famille*, p. 27–43. Armand Colin.

STANCZAK K. & AUGENSTEIN I. (2021). A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

UCHIDA S. (2024). Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations. *Applied Corpus Linguistics*, **4**(1), 100089. DOI : https://doi.org/10.1016/j.acorp.2024.100089.

WANNA S., SOLOVYEV N., BARRON R., EREN M. E., BHATTARAI M., RASMUSSEN K. O. & ALEXANDROV B. S. (2024). TopicTag: Automatic Annotation of NMF Topic Models Using Chain of Thought and Prompt Tuning with LLMs. In *Proceedings of the ACM Symposium on Document Engineering 2024*, DocEng '24, New York, NY, USA: Association for Computing Machinery. DOI : 10.1145/3685650.3685667.

WAUQUIER M., FABRE C. & HATHOUT N. (2020). Semantic discrimination of technicality in French nominalizations. *Zeitschrift für Wortbildung / Journal of Word Formation*, **2020**(2), 100–121.

A Listes des thématiques

A.1 Annotation des voisins distributionnels

Thématique	Exemples de voisins
âge	gosse, octogénaire
argent	banquière, rente
armée	adjudant, garnison
art	poète, mécène
caractéristique	blonde, crapule
crime/délit	assassin, comploter
éducation	élève, mentor
enfance	gamine, doudou
foyer	domestique, maisonnée
héritage	cohéritier, testamentaire
lien/relation	collègue, soupirant
mort	enterrement, suicider
métier/occupation	actrice, rentier
notable/noble	comtesse, tsar
procréation	naissance, allaiter
religion	baptême, adoratrice
santé	docteur, hypocondriaque
sentiment/amour	chéri, jalousie
sexe	prostitué, virginité
titre/fonction	chevalier, ursuline

A.2 Etiquettes des topics par ChatGPT

Étiquette	Exemples de mots
Arts & Culture	scène, musée, épisode
Célébrations & Rites	mariage, anniversaire, fête
Conflits & Drames	tuer, accident, guerre
Économie & Argent	entreprise, compagnie, associer
Éducation & Transmission	école, étude, professeur
Famille & Origines	épouser, famille, naître
Histoire & Héritage	hériter, château, siècle
Naissance & Mort	naître, mourir, fœtus
Nature & Territoire	île, mètre, design
Personnalités & Figures historiques	peintre, écrivain, français
Pouvoir & Société	roi, président, conseiller
Quotidien & Environnement	maison, lit, vivre
Relation & Sentiments	amie, amour, jeune
Santé & Vieillissement	santé, vieillesse, maladie
Spiritualité & Croyances	évêque, prophète, église
Sports & Loisirs	olympique, prix, joueur
Travail & Professions	carrière, entreprise, militaire

Exemple de prompt pour la génération d'étiquettes à partir des topics :

J'ai lancé un processus de topic modeling sur un extrait de la Wikipédia, en ne retenant que des phrases qui contiennent un mot relatif à la famille (frère, tante, cousin etc.). J'ai extrait un certain nombre de topics dont tu trouveras ci-dessous les mots les plus significatifs (numéro du topic + poids + liste des mots). Tu dois me donner un court descriptif pour chaque topic, en utilisant un terme générique et court (e.g. "nourriture", "sexe", "sentiment" etc.). Tu es encouragé à utiliser le même terme pour plusieurs topics qui te semblent similaires. Si tu ne vois aucun point commun entre les mots, utilise le mot "BRUIT". Donne-moi juste le numéro de topic suivi du caractère tabulation et de ta proposition rédigée en minuscules.

Exemple de prompt pour le regroupement d'étiquettes :
Je vais te donner une liste de 67 thématiques qui sont liées aux relations familiales. Peux-tu les regrouper en 15 à 20 catégories plus générales. Tu donneras la liste des catégories générales suivie d'une tabulation suivie des thématiques que tu as regroupées séparées par des tabulations. Tu peux créer une thématique DIV si tu n'arrives pas à regrouper certaines thématiques.

Exemple de prompt pour l'assignation finale d'étiquettes aux topics :
J'ai lancé un processus de topic modeling sur un extrait de la Wikipédia, en ne retenant que des phrases qui contiennent un mot relatif à la famille (frère, tante, cousin etc.). J'ai extrait 50 topics dont tu trouveras ci-dessous les mots les plus significatifs (numéro du topic + poids + liste des mots). Tu dois me donner un court descriptif pour chaque topic, en utilisant une des 17 catégories que je te soumets. Chaque catégorie est suivie entre parenthèses de catégories plus spécifiques qui précisent leur sens. Tu es encouragé à utiliser la même catégorie pour plusieurs topics qui te semblent similaires. Si tu ne vois pas ou très peu de point commun entre les mots, utilise le mot "BRUIT". Donne-moi juste le numéro de topic suivi du caractère deux points (:) et de ta proposition.

A.3 Thématiques proposées par ChatGPT

ChatGPT-1	ChatGPT-2	ChatGPT-3	ChatGPT-4
autorité	Alliance	Affection	Culture
cinéma	Famille élargie	Alliance	Droit
droit	Fratrie	Ancêtres	Généalogie
éducation	Génération	Autorité	Héritage
famille élargie	Histoire et société	Collatéraux	Histoire
fratrie	Légalité	Coutumes sociales	Littérature
généalogie	Lien biologique	Fratrie	Mariage
histoire	Parentalité	Générations	Mythologie
littérature	Sexe	Genre	Noblesse
mariage	Terminologie culturelle	Juridique	Parenté
mythologie	Transmission patrimoniale	Lien direct	Religion
parenté biologique	Usages religieux	Transmission	Sociologie
parenté légale			
relations intergénérationnelles			
relations sociales			
statut familial			

Exemple de prompt - modèle ChatGPT-4 *Je vais te proposer 25 mots qui désignent des relations familiales. Identifie pour chacun de ces mots des thématiques qui les caractérisent. Au total tu dois me proposer au moins 12 thématiques. Ces thématiques sont liées à l'usage de ces mots dans les articles de Wikipedia. Produis un fichier csv téléchargeable avec en ligne les 25 mots et en colonne les thématiques, avec un score entre 0 et 10 qui donne le degré de caractérisation du mot par la thématique. Voici les 25 mots : beau-frère, beau-père, belle-fille, belle-mère, belle-sœur, cousin, cousine, épouse, époux, fille, fils, frère, gendre, grand-mère, grand-père, mari, mère, neveu, nièce, oncle, père, petit-fils, petite-fille, sœur, tante.*

B Analyses en composantes principales

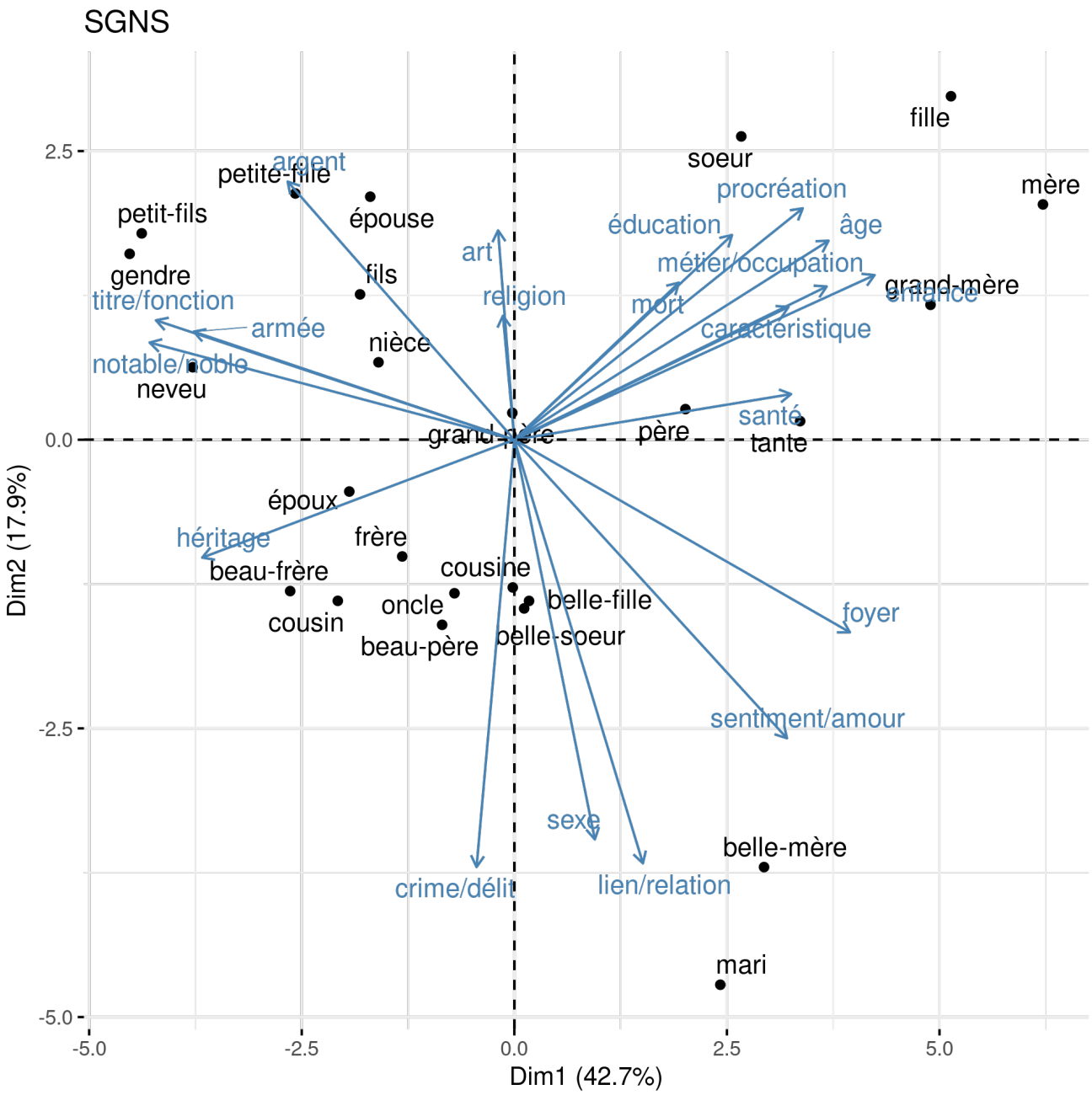

FIGURE 2 – Analyse en composantes principales : thématiques associées aux amorces - modèle SGNS, catégories manuelles

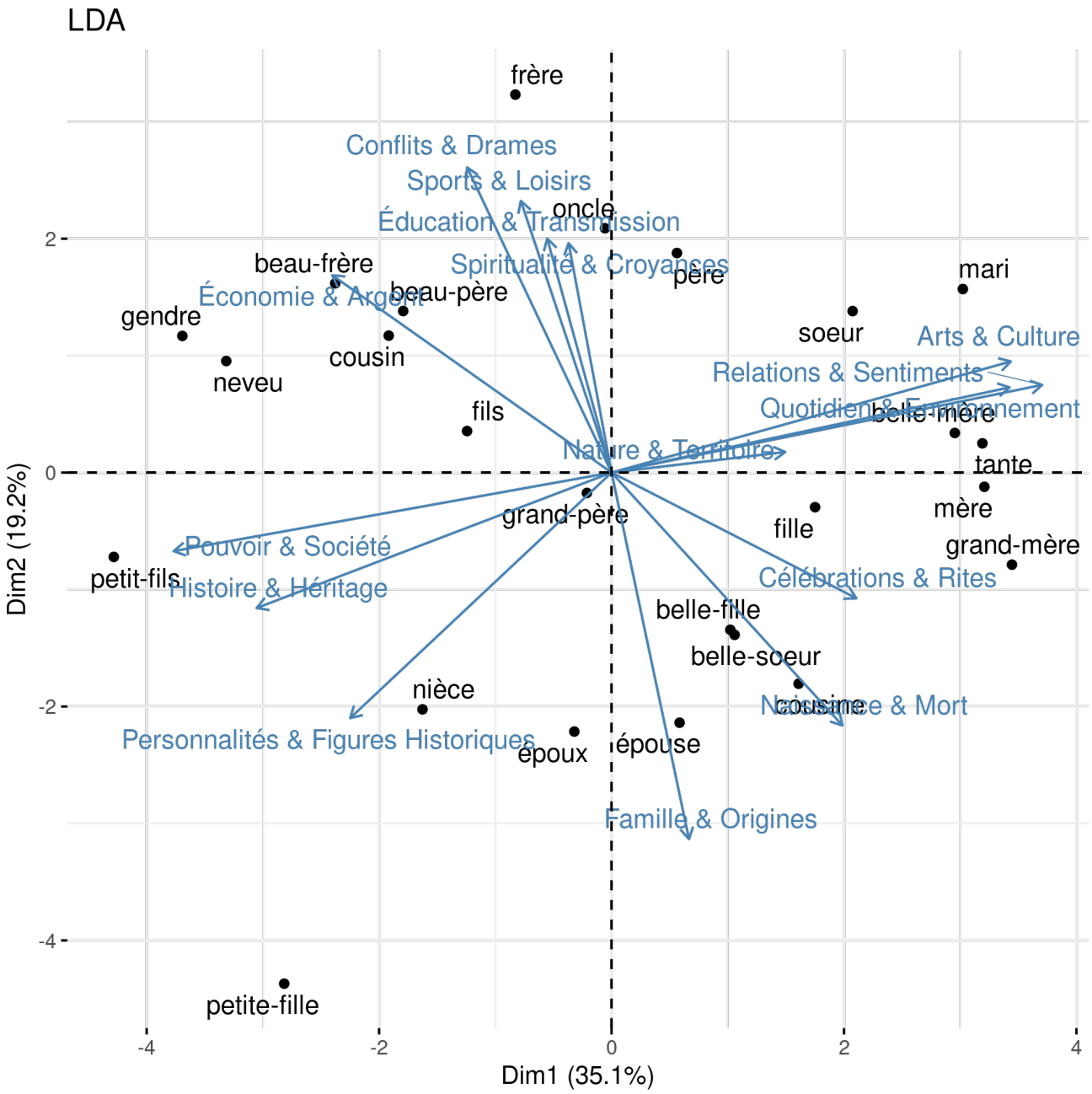

FIGURE 3 – Analyse en composantes principales : thématiques associées aux amorces - modèle LDA, catégories proposées par ChatGPT