Ecole Nationale Supérieure des Mines de Paris

Travail d'option

effectué à

l'ATILF (Nancy, CNRS)

Analyse et modélisation sémantiques à partir de ressources lexico-sémantiques

---

Reutenauer Coralie

Année 2008

Remerciements

Mes remerciements s'adressent avant tout à mes quatre tuteurs. Merci infiniment à Evelyne Jacquey qui m'a consacré temps et énergie sans compter, dont les conseils éclairés m'ont permis de progresser à chaque nouvelle étape et dont le dynamisme inépuisable a su me stimuler en permanence. Merci à Mathieu Valette qui m'a fait découvrir les terres inconnues de la sémantique, m'a permis d'avoir des échanges extrêmement enrichissants et qui a toujours su être attentif à mes interrogations et à ma progression. Merci à Jean-Marie Pierrel, pour l'intérêt qu'il a manifesté pour mon projet tout au long de mon stage et pour m'avoir ouvert non seulement les portes de son laboratoire, mais aussi celles d'un univers extraordinaire, et qui a ainsi réussi à me communiquer la passion de la recherche. Merci à Pierre Chauvet, sans lequel je n'aurais pu faire ce stage, tant pour les efforts qu'il a déployés afin de satisfaire mes demandes que pour son extrême disponibilité, l'ouverture dont il a fait preuve et tout ce qu'il a mis en œuvre pour m'accompagner dans ma progression.

Je tiens aussi à remercier Etienne Petitjean qui m'a permis de faire d'immenses progrès en Java et sans lequel mon programme n'aurait probablement pas fonctionné ; Mick Grzesitchak, qui a su m'apporter son secours sur nombre de questions informatiques et m'a initiée à Sémy ; Bertrand Gaiffe, qui a su m'éclairer à plus d'une occasion et a fait avancer mes réflexions par ses remarques pertinentes ; Sandrine Ollinger pour sa présence, son aide sur des points problématiques et l'intérêt qu'elle a manifesté pour mon travail.

Je souhaite enfin exprimer ma reconnaissance à l'ATILF et ses membres qui ont su si bien m'accueillir et, plus largement, à tous ceux qui se sont intéressés à mon projet et m'ont soutenue pour le mener à bien.

Table des matières

I) Objectifs : bâtir un modèle intégrant des éléments d'une sémantique de corpus

Les langues naturelles, outils quotidiens de communication, apparaissent à plusieurs égards comme un ensemble structuré, avec ses règles syntaxiques ou encore ses régularités morphologiques. Le sens qu'elles véhiculent constitue, de mon point de vue, une de leurs raisons d'être majeure. La sémantique, discipline de la linguistique qui a pour objet l'étude du sens, s'attaque donc à un aspect fondamental du langage et constituera le socle de mon stage.

Le sens est, par essence, subjectif. L'interprétation d'un texte, le texte constituant une forme particulière de l'usage de langues naturelles, varie selon les individus, l'époque et, de manière plus générale, l'environnement de ce texte. Pourtant, nous nous comprenons. Ce consensus pourrait donc être considéré comme le reflet d'une convergence sémantique. Les linguistes tentent de comprendre les mécanismes qui régissent cette convergence et d'aller au-delà de leur intuition linguistique. Pour ce faire, ils collectent des données réelles. Cependant, la masse de données générées est très importante et, pour la synthétiser et l'exploiter, le recours à une modélisation mathématique des phénomènes de convergence et de variation sémantiques en présence s'avère nécessaire.

Cette modélisation, un des enjeux majeurs du stage effectué, met en jeu des disciplines jeunes à la croisée de la linguistique, de l'informatique et des mathématiques : le traitement automatique des langues, la statistique linguistique et la sémantique textuelle. Ces sciences abordent le fonctionnement des langues naturelles, en particulier leur fonctionnement sémantique, à différents niveaux : le niveau lexical, c'est-à-dire le niveau du mot, supra-lexical, fondé sur des unités langagières plus étendues que le mot (phrase, paragraphe, texte) et infra-lexical, s'appuyant sur des unités plus petites que le mot. Ce dernier niveau, niveau central de mon étude, repose sur le principe suivant : à tout mot peut être affecté un ensemble d'unités de sens minimales, appelées traits sémantiques ou sèmes. Les traits sémantiques qui composent un mot peuvent être partagés par d'autres mots, comme par exemple l'idée de mouvement ou encore l'opposition entre concret et abstrait. Ils interagissent et déterminent ainsi notre perception du sens.

Par ailleurs, mon travail se fonde sur l'hypothèse suivante : le sens n'est pas intrinsèque mais dépend de son environnement, environnement que nous appellerons le « cotexte » dans le cadre d'un texte donné. Ainsi, les relations entre traits sémantiques et l'émergence du sens reposent sur l'usage. Le sens n'est pas figé comme dans les ressources encyclopédiques, il est vivant, mouvant et évolue dans le temps. Il est dépendant des situations, des interlocuteurs, tout comme les usages[1].

Cette approche linguistique du sens, représentée notamment par la sémantique interprétative ou textuelle, permet d'étudier les tendances sémantiques globales d'un texte, mais aussi les variations fines à plus petite échelle, notamment par rapport à un mot, c'est-à-dire au niveau lexical. Les recherches développées au cours de mon stage et présentées dans ce rapport se centrent sur ce dernier point : le mot et l'étude de ses variations locales en étudiant les variations au niveau infra-lexical par l'intermédiaire des traits sémantiques ou sèmes.

Pour mesurer les déplacements sémantiques, je me suis efforcée de rechercher des modèles pertinents dans des domaines semblables au mien, d'adapter ces modèles et d'étudier leur qualité. Cette qualité a été estimée à l'aune de mes objectifs, à savoir obtenir une représentation mathématique globale du contenu sémantique d'un texte ou corpus de textes et observer les variations du contenu sémantique d'un mot en un point du texte.

[1] Nous rappelons cependant que la présente étude s'intéresse essentiellement à l'influence des usages représentés par les textes. En effet, les linguistes n'ont pas les outils théoriques nécessaires à la modélisation des situations comme celle des interlocuteurs en tant qu'individus.

II) Cadre général : l'étude des langues naturelles, en particulier du français

2.1) Le Traitement Automatique des langues

Avec l'émergence des NTIC et la nécessité de gérer l'information, l'ingénierie des langues a pris une dimension majeure, dont les enjeux et avancées sont décrits dans [Pierrel, 1997]. Elle s'est en particulier concrétisé à travers le TAL, Traitement Automatique des Langues, aussi appelé TALN (Traitement Automatique du Langage Naturel),

Le TAL est né vers le milieu du XXe siècle aux Etats-Unis. Il a pour objet le traitement automatique à partir d'outils informatiques, linguistiques et formels de données textuelles (textes écrits ou oraux ou encore unités linguistiques).

Comme le soulignent [Cori & Léon, 2002], les frontières du TAL ne sont pas clairement définies. Il balance entre science et technologie, oscille entre visées théoriques et industrielles. Sa délimitation est donc délicate. Quelques éléments permettent cependant de saisir globalement ce qu'il représente.

Le TAL repose sur quatre disciplines principales : la linguistique, l'informatique, les mathématiques et les sciences cognitives. Né dans une optique de traduction automatique, il voit son champ d'investigation s'étendre rapidement pour recouvrir des domaines très variés. Selon [Miller & Torris, 1990] cité par [Cori & Léon, 2002], il s'intéresse à la linguistique théorique, qu'il cherche à décrire explicitement ; à l'informatique théorique pour l'optimisation des algorithmes et programmes mis en place ; à « l'étude mathématique des propriétés formelles des outils de traitement et théories linguistiques » ([Miller & Torris, 1990], p.15) ; à l'intelligence artificielle et aux théories cognitives.

Sur le plan linguistique, il se situe à différents niveaux d'observation : le niveau morpho-lexical, qui s'attache à l'étude de la structure des mots (morphologie) et à la classification et au recensement des formes d'une langue (lexicologie) ; le niveau syntaxique (par exemple, pour les grammaires d'une langue) ; sémantique (étude du sens) ; pragmatique (contextualisation). Ces différentes approches sont complémentaires, souvent imbriquées, comme par exemple les démarches s'intéressant à l'interface syntaxe / sémantique.

L'existence du TAL se justifie par deux raisons principales : il permet d'une part d'analyser de grands corpus de textes et d'autre part de mettre en place et analyser des modèles formels.

Dans le vaste champ d'investigation du TAL, mon travail se positionne au niveau sémantique.

Le TAL a de nombreux domaines d'application : la recherche d'information, la traduction automatique, la classification de textes, le filtrage d'information, la correction automatique, la génération automatique de textes (résumé par exemple) ou encore la compréhension automatique des textes. Les domaines d'application de mon sujet sont principalement la recherche d'information et la classification de textes.

2.2) Etablissement d'accueil : l'ATILF

Le laboratoire ATILF (Analyse et Traitement de la Langue Française) au sein duquel j'ai effectué mon stage est une unité mixte de recherche du CNRS (département Homme et Société) et de Nancy Université, Campus Lettres et Sciences Humaines et Université Henri Poincaré. Il est issu du rapprochement de l'INALF (Institut National de la Langue Française) et de l'équipe d'accueil de l'université LANDISCO (Langue, discours, cognition – université Nancy 2). Ses champs d'investigation se situent à la croisée de différentes disciplines : linguistique, informatique et mathématiques.

Le projet phare de l'ATILF est le Trésor de la Langue Française informatisé (TLFi). Le Trésor de la Langue Française (TLF), dont le TLFi est la version informatisée, est un dictionnaire de langue française des XIXe et XXe siècles en 16 volumes et un supplément. TLF et TLFi sont le fruit d'un travail de plus de quarante ans, débuté sous la direction de Paul Imbs en 1957. La version actuelle du TLFi est disponible sur le web en accès libre (site http://www.atilf.fr/tlfi.htm) et sur CD-Rom. Ce

dictionnaire informatisé se distingue par les fonctionnalités de recherche qu'il propose : recherche simple avec affichage de l'article et outils de visualisation des différents éléments de l'article (définition, exemples, …), recherche assistée et requêtes complexes.

L'ATILF ne se limite pas au TLFi : d'autres projets d'envergure ont été menés. Citons la réalisation de Frantext, base textuelle constituée de près de 4000 textes littéraires français d'environ 1000 auteurs du XVIe au XXIe siècle. Les textes peuvent y être consultés par recherches simples ou complexes. Une version partielle de Frantext est accessible librement à l'adresse http://www.atilf.fr/frantext.htm. Mentionnons également les nombreuses études portant sur l'ancien français et ayant conduit au DMF (Dictionnaire du Moyen Français, accessible en ligne à partir de l'adresse www.atilf.fr/dmf), autre produit phare de l'ATILF contenant près de 120000 articles sur la langue française de 1330 à 1500. Enfin, le Französisches Etymologisches Wörterbuch (FEW), dictionnaire étymologique du moyen français, offre une approche approfondie du galloroman, avec une description du gascon, de l'occitan, du francoprovençal et de dialectes d'oïl. Celle-ci est étayée par toutes les données accumulées de la lexicographie française et recense les évolutions morphologiques et sémantiques du galloroman au cours des siècles.

III) Vers la modélisation : cadre théorique, ressources et outils disponibles

3.1) Théorie linguistique : la sémantique interprétative ou sémantique textuelle

La sémantique interprétative, développée à partir des années 80 par François Rastier (1987, 1991, 2001), est une théorie unifiée visant à décrire tous les paliers de la textualité, du mot au texte, à partir des mêmes outils conceptuels. Parmi ceux-là, le *sème* (ou trait sémantique), hérité de la tradition structuraliste (Saussure, Greimas, Pottier), présente un intérêt tout particulier pour notre propos.

3.1.1 Une sémantique des pratiques

Deux traditions fondent la sémantique d'aujourd'hui : la tradition rhétorique-herméneutique qui traite de textes et la tradition logico-grammaticale. Cette seconde approche, courant dominant dans la communauté linguistique, a construit la sémantique sur de petites échelles : étude du sens au niveau du mot ou encore de la phrase. L'apport des cotextes et contextes y a alors été sous-estimé et négligé. Par cotexte, nous entendons l'ensemble des unités sémantiques qui ont une influence sur une unité donnée et sur lequel elle-même a une incidence. Le contexte renvoie à l'environnement extralinguistique. Les textes et, dans une certaine mesure, les contextes matérialisent la notion d'usage. La sémantique interprétative s'attache à l'étude d'un sens non pas ontologique, c'est-à-dire d'un sens par essence, intrinsèque au mot, mais d'une variété de sens associée aux textes et aux usages.

L'environnement (textes, contextes,…) influence le sens des mots sur plusieurs plans. Le genre est un premier cadre d'influence. En effet, celui-ci met en jeu un univers sémantique dans lequel les unités de sens mobilisées s'inscriront. Par exemple, les unités de sens activées pour le mot « essence » seront, dans le cas général, plutôt reliées au pétrole et à des notions économiques dans un corpus journalistique, alors qu'elles feront écho à l'être et à l'existence dans des traités de philosophie. L'époque a également un impact sémantique : les pratiques sociales changent au cours des siècles, ainsi que le sens des mots. Considérons le groupe nominal « le mari déçu » : dans des conversations du XXIe siècle, on imaginera plutôt un échec de l'épouse sur un terrain quelconque (championne sportive détrônée, rôle de représentation mal tenu, …) et le mari affecté par l'incapacité de sa femme à satisfaire ses attentes ; dans du Molière, cette expression évoquera le mari trompé par sa femme. De plus, la sémantique interprétative fait l'hypothèse que les discours et genres textuels reflètent le cadre socio-culturel, ce qui influera également sur les unités de sens activées. Enfin, la taille des cotextes joue aussi un rôle important : les unités de sens émergentes ne seront pas toujours les mêmes si on se borne à une phrase, un paragraphe, qu'on s'étend à un chapitre, un texte ou encore à un corpus de textes. Un concept bien connu illustrant cet aspect est celui de l'intertextualité : celle-ci ne peut être activée que si le lecteur se place non pas simplement au niveau du texte qu'il lit mais se place dans un univers sémantique constitué de lectures antérieures.

Une notion-clé s'inscrit dans ce cadre de l'usage : le parcours interprétatif. Rastier, dans son glossaire repris par [Missire, 2006], définit le parcours interprétatif comme une « suite d'opérations permettant d'assigner un ou plusieurs sens à un passage ou à un texte ». En clair, cela signifie que chacun construira sa propre approche du sens selon différents paramètres : son milieu d'origine, son époque, sa culture, le moment et la situation dans laquelle il est confronté au texte,… Ainsi, l'interprétation est influencée par de multiples paramètres, variables selon les individus.

Enfin, soulignons que l'approche textuelle prend le contre-pied des références dictionnairiques ou encyclopédiques. Au mot au sens figé par ces ressources s'oppose un mot au sens évolutif selon les contextes (situation d'énonciation ou de production du texte). Les cotextes liés à ces situations permettent alors non seulement de désambiguïser un mot polysémique mais aussi d'introduire des variations sémantiques pour un mot monosémique.

3.1.2 Formalisation de cette théorie : les traits sémantiques ou sèmes

L'introduction d'entités particulières, les traits sémantiques ou sèmes, a permis de formaliser les principes exposés ci-dessus. Cette démarche s'inscrit dans un cadre infra-lexical qui considère que les mots sont décomposables en unités de sens plus petites. Les traits sémantiques ou sèmes constituent les unités de sens minimales. Chaque mot comprend un sémème, ensemble structuré de traits sémantiques.

Les traits sémantiques d'un sémème peuvent être classés en différentes catégories. Ainsi, les sèmes peuvent être génériques ou spécifiques. Des sèmes génériques sont des sèmes qui indiquent l'appartenance à une classe, une famille plus vaste, comme le domaine auquel le mot appartient. Au contraire, les sèmes spécifiques sont les sèmes permettant de distinguer le mot par rapport aux autres mots des mêmes domaines ou classes. Par exemple, pour le mot poirier, le trait sémantique /arbre/ est un sème générique qui ramène à une famille plus vaste ; le trait sémantique /poire/ est en revanche un sème spécifique, propre au mot poirier.

Ces différentes catégories de traits ont leur importance dans le cadre mathématique où nous nous plaçons. Ainsi, de bonnes mesures de distance entre mots devraient refléter la structure en traits génériques et spécifiques. Les sèmes génériques seraient facteurs de rapprochement sémantique entre deux mots et les sèmes spécifiques facteurs d'éloignement. L'analyse linguistique, avec répartition des sèmes en sèmes génériques et spécifiques, ouvre des perspectives sur le mode de validation d'un modèle mathématique.

Par ailleurs, deux statuts peuvent être affectés aux sèmes : le statut de sème inhérent et celui de sème afférent. Un sème est dit inhérent s'il est hérité d'un mot, par exemple le sème /noir/ pour *corbeau*. Il est au contraire dit afférent s'il est greffé à un mot du fait d'un cotexte particulier. Par exemple, *cheval* aura pour sème afférent /jouet/ dans l'expression *cheval de bois.* Cette notion de sèmes afférent et inhérent soulève un problème majeur d'une modélisation idéale : celle-ci doit considérer un mot non comme un ensemble structuré de taille fixe, susceptible d'évoluer uniquement au niveau de sa structure interne, mais comme un ensemble de taille variable, auquel peuvent être ajoutés des éléments quelconques de l'univers (espace constitué de l'ensemble des points ; dans notre cas, il s'agirait de l'ensemble des traits sémantiques de la langue française).

Ajoutons aux notions abordées celle de forme sémantique et de molécule sémique. Une forme sémantique est un groupement stable de sèmes spécifiques articulés par des relations structurales. Une molécule sémique est un cas particulier de forme sémantique. Je n'approfondirai pas les différences entre forme sémantique et molécule sémique, approche détaillée qui sort de mon champ de compétences et m'écarte de l'objet de ce travail de stage à l'intérieur duquel il m'est actuellement possible d'utiliser indifféremment l'un pour l'autre, approximation que le lecteur voudra bien me pardonner. Ce concept de molécule sémique m'a paru important car il implique la structuration du sémème et fait écho à la notion de clusterisation en mathématiques. Il ouvre donc des pistes de réflexion intéressantes sur la modélisation.

3.1.3 Phénomènes observés

Les traits sémantiques sont soumis à divers phénomènes en contexte, illustrés dans [Valette, 2004] et [Valette & Grabar, 2004] : l'activation, la virtualisation, la domanialisation et dédomanialisation ; ils peuvent se regrouper en noyau sémique ou être à l'origine d'isotopies et enrichir le sémème d'un mot.

Tout d'abord, les traits sémantiques peuvent être activés ou au contraire virtualisés, c'est-à-dire inhibés en contexte. Par exemple, dans l'expression *Un verre de rouge*, le trait sémantique /alcool/ est activé dans le mot *rouge*. En revanche, dans *un chat siamois*, le trait sémantique /jumeaux/ du mot siamois est inhibé.

Un autre phénomène est celui de la domanialisation : le sens d'un un mot peut se voir rattaché à un domaine particulier dans un contexte donné. A l'inverse, un mot peut être dédomanialisé, c'est-à-dire

qu'un sème générique qui le caractérise peut être inhibé en contexte. Les deux exemples cités ci-dessous, tirés de [Valette & Rastier, 2008], illustrent ces notions de domanialisation et dédomanialisation.

Considérons le premier exemple :

> Si l'on devient de plus en plus riche, on remplacera peut-être progressivement le McDo quotidien par des toasts au caviar, du homard, des omelettes aux truffes blanches et d'autres choses encore plus appétissantes et aussi raffinées que coûteuses (Forum du site *teleologie.org*, 3.03.2001)

Le mot cavíar connaît dans cet exemple une domanialisation gastronomique (présence d'un trait /gastronomie/). Inversement, dans la presse sportive, l'utilisation de cavíar pour qualifier une belle passe est l'illustration même, sur le plan sémantique, d'une dédomanialisation accompagnée d'une redomanialisation. Le domaine dans lequel cavíar se situe n'est plus la gastronomie ou le luxe mais le football (allocation d'un trait sémantique /sport/).

L'isotopie recouvre une réalité assez différente des deux précédentes. Il s'agit d'un effet de la récurrence d'un sème qui se traduit par la présence répétée dudit sème dans un texte à intervalles réguliers. Le mot correspondant à l'unité de sens isotopique peut tout à fait être rare dans le texte, voire absente.

Par ailleurs, dans les textes, on peut assister à des regroupements de traits sémantiques plus ou moins variables. Ces regroupements varient mais semblent parfois présenter des éléments communs récurrents : le noyau sémique. Soulignons que l'existence d'un noyau sémique, pour l'ensemble des mots ou, plus vraisemblablement, certaines catégories de mots, n'a pas encore été démontrée. Une modélisation du phénomène et des études statistiques des résultats permettrait de répondre, ou de formuler un début de réponse à cette question ouverte. Ces questions de regroupements sémantiques ont au demeurant déjà fait l'objet d'études (voir [Valette, Estacio-Moreno, Petitjean & Jacquey, 2006]).

Enfin, un dernier phénomène à mentionner est l'enrichissement du sémème. Cette question d'enrichissement n'est pertinente que si l'on considère que le sémème d'un mot est, à un instant donné, un ensemble fini de traits sémantiques structurés. Ce sémème constitue le sémème de référence. L'étude du mot dans une série de cotextes peut faire émerger que certains traits sémantiques manquent dans le sémème de référence, tandis que d'autres peuvent sembler présents à tort. Le sémème de référence peut alors être enrichi ou appauvri. Le nouveau sémème peut alors être considéré comme une nouvelle représentation du mot relative à une classe de cotextes possédant des caractéristiques communes. En réitérant l'étude sur plusieurs classes de cotextes, il sera possible de générer des sémèmes profilés en fonction d'usages (ceux représentés par la classe de cotextes choisis). L'enrichissement met donc en lumière des mécanismes fins qu'une approche mathématique pourrait aider à normaliser.

Dans mes démarches, je me suis efforcée de faire émerger par des méthodes mathématiques certains de ces phénomènes ou états : l'activation et l'inhibition de traits sémantiques ; l'existence ou non d'un noyau sémique ; la structuration en molécules sémiques dans un contexte donné. Pour des questions de temps, je n'ai pu approfondir la question de l'enrichissement du sémème. L'étude de l'isotopie, envisagée dans un premier temps, a été écartée puisqu'elle ne s'inscrivait pas dans la démarche consistant à aller du global (texte, corpus) au local (mot).

3.2) Théories mathématiques pour l'analyse linguistique

L'analyse linguistique pourrait, certes, être exclusivement du ressort des linguistes qui possèdent à la fois la connaissance des mécanismes de langue et une intuition qui semble difficilement quantifiable. La question de la modélisation est d'ailleurs très débattue : certains soutiennent qu'elle est impossible. Si le recours à des métriques n'est peut-être pas à même de traiter finement toutes les subtilités du langage, il peut cependant faire émerger des tendances, mettre à jour des mécanismes caractérisés par

certaines régularités. De plus, il ouvre des perspectives sur le traitement et l'analyse de grandes masses de données (celles de corpus par exemple), opération qui dépasse les capacités humaines.

Différents modèles et procédés mathématiques ont retenu mon attention, depuis [Muller, 1968] ou [Habert & Nazarenko, 1997] à [Victorri, 2005 & 1994], [Venant, 2004] ou [Landauer, Foltz & Laham, 1998] : ils paraissaient robustes, transposables au moins sur certains plans et certains de leurs rouages appropriés aux outils que je souhaitais développer pour les analyses sémantiques envisagées.

3.2.1) De la statistique linguistique à tf-idf

La plupart des supports mathématiques développés en linguistique puisent leur source dans une science fondée en France dans les années soixante par Charles Muller (cf [Muller, 1968]) : la statistique linguistique.

Cette discipline développe les outils d'analyse de la linguistique. Elle étudie par exemple la structure et l'étendue du vocabulaire, la pertinence de distributions en fréquence de certains mots ou catégories grammaticales, etc. Elle utilise divers outils statistiques : indicateurs moyens, indicateurs de dispersion, coefficient de corrélation, tests statistiques types. Un test statistique utilisé avec succès dans diverses études est celui du χ^2, dont on trouvera les détails dans [Hatchuel & Tonneau, 1996]. Celui-ci s'effectue de la manière suivante :

Soit un tableau constitué de m lignes et p colonnes. Le nombre de degrés de liberté est de $(n-1)\times(p-1)$. On note n_{ij}, $1 \le i \le m, 1 \le j \le p$, $n_{i\bullet}$ la somme des coefficients de la ligne i, $n_{\bullet j}$ la somme des coefficients de la ligne j, n la somme totale des coefficients.

	$\vdots$		
$\dots$	n_{ij}	$\dots$	$n_{\bullet j}$
	$\vdots$		
	$n_{i\bullet}$		n

L'étape suivante consiste à établir une valeur théorique moyenne m_{ij} sous l'hypothèse d'indépendance des lignes et colonnes pour chaque coefficient : $m_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$. On calcule ensuite la valeur observée de la variable de test : $\chi^2 = \sum_{1 \le i \le m, 1 \le j \le p} \frac{\left(n_{ij} - m_{ij}\right)^2}{n}$.

Cette valeur est ensuite comparée avec un tableau de distribution du χ^2 tabulé en degrés de liberté. Chaque colonne correspond au seuil de probabilité au-delà duquel l'hypothèse de départ (dans notre cadre, hypothèse d'équirépartition ou encore de répartition non significative des occurrences ou cooccurrences) est rejetée. Cette méthode est intéressante, bien que le stade consistant à sommer les écarts au carré entre valeurs réelles et valeurs théoriques fasse perdre l'information apportée par chaque coefficient. Elle est la source de certaines transformations que j'ai effectuées.

[Muller, 1968] propose une synthèse des connaissances accumulées sur les lois lexicales existantes et, à travers des études statistiques, discute de leur validité. La loi de Zipf a particulièrement retenu mon attention. Considérons les mots d'un texte classés par ordre de fréquence décroissant. Soit n le rang d'un mot, $f(n)$ sa fréquence. La loi de Zipf est, d'après [Lemire 2008] et [Muller, 1968], de la forme $f(n) = \frac{K}{n}$, K constante. Cette loi a par la suite été généralisée par Mandelbrot. La loi dite de Zipf-Mandelbrot est de la forme $f(n) = \frac{K}{(a+bn)^c}$, où a, b, c et K sont des constantes. Cette loi, dite loi empirique, reflète une tendance générale du lexique, affirmation étayée par de nombreuses études.

Au niveau des phénomènes linguistiques en jeu, elle indique que le comportement général de la distribution des occurrences n'est pas uniforme : il existe un petit nombre de mots très fréquents et un grand nombre de mots très rares.

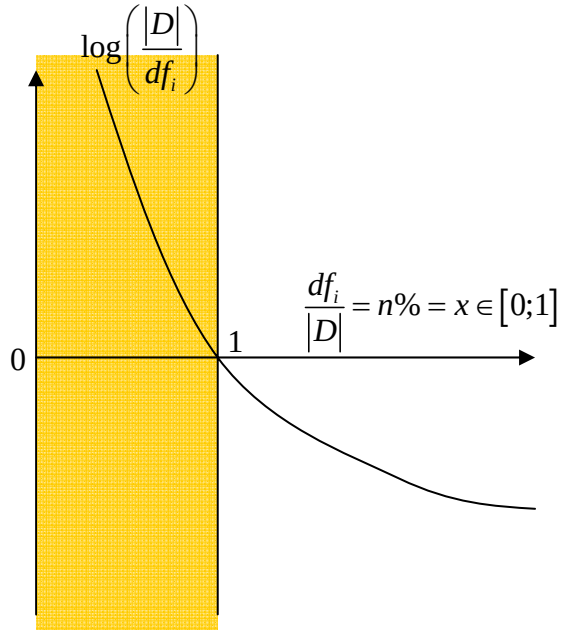
De cette loi découle une méthode mathématique assez utilisée en recherche d'information : la méthode tf-idf. Elle s'appuie également sur deux autres constatations. Première constatation : les mots très présents, statistiquement présents dans une forte proportion de documents constituant un corpus, ne sont pas discriminants. Deuxième constatation : les mots les plus fréquents n'apportent que peu d'information sémantique.

La méthode tf-idf s'appuie sur le nombre d'occurrences ou sur la fréquence d'un mot (tf : term frequency) et sur la distribution de ce mot dans différents textes, paragraphes ou autres unités textuelles (idf : inverse document frequency) Considérons un corpus constitué de documents. Soit $|D|$ le nombre de documents du corpus et df_i le nombre de documents contenant le mot i. $\frac{df_i}{D}$ correspond donc à la proportion de documents contenant le mot i.

$$idf(i) = -\ln\left(\frac{df_i}{|D|}\right)$$

Notons l'utilisation du logarithme. Celle-ci trouve ses sources dans la théorie de l'information : la quantité d'information I relative un événement e_i ayant la probabilité d'occurrences pi est : $I(e_i) = -\log_2(p_i)$ (voir [Rouchaleau, 2008] p.17). La probabilité est ici remplacée par la fréquence d'apparition (en termes de présence / absence, sans décompte multiple des occurrences) dans les documents.

La fonction idf permet de représenter le poids du mot dans le corpus. Elle accorde un poids important aux termes rares et un poids faible aux termes très fréquents, avec une décroissance du poids de plus en plus lente.



$tf(i, j)$ = fréquence du mot i dans le document j.

Cette fonction représente le poids du mot à l'intérieur d'un document. Ce poids croît proportionnellement au nombre d'occurrences du mot.

La formule générale de tf-idf définit le coefficient suivant : $tfidf(i, j) = tf(i, j) \times idf(i)$. Ce coefficient peut s'interpréter comme un coefficient de significativité : les termes qui ont une forte

significativité pour un document donné sont très présents dans ce document, mais rares dans les autres documents. Des termes très présents dans le document considéré mais également dans tous les autres documents ont une significativité relativement faible (termes non discriminants, donc n'apportant que peu de valeur ajoutée).

La transformation tf-idf, considérée comme une référence par les scientifiques en linguistique, a été retenue pour certaines des expériences menées au cours de ce stage. Insistons cependant sur un point : la loi de Zipf ainsi que la transformation tf-idf ont été mises en place et testées au niveau lexical, c'est-à-dire au niveau des mots. Les études au niveau infra-lexical, à l'aide des traits sémantiques dans le cadre de ce travail, sont récentes et encore au stade exploratoire. . Il n'est donc pas à écarter que les résultats des mêmes lois doivent être interprétés un peu différemment au niveau infra-lexical.

3.2.2) Modèles récents : métriques et distances sémantiques

3.2.2.1) Modélisation de polysémie lexicale par Bernard Victorri

[Victorri, 2005] développe un modèle pour désambiguïser le sens d'un mot ayant plusieurs sens possibles, c'est-à-dire trouver le sens approprié d'un terme polysémique, à partir d'un dictionnaire de synonymes constitué au préalable. La méthode utilisée comporte deux étapes.

La première étape se déroule comme suit : choix d'un adjectif à désambiguïser ; détermination de la liste des synonymes de cet adjectif ; constitution de cliques, c'est-à-dire de regroupements ou clusters de synonymes à partir du dictionnaire de synonymes ; calcul de distances entre les cliques à partir d'une matrice valuée en fonction de la présence (valeur 1) ou de l'absence (valeur 0) d'un synonyme dans une clique.

La deuxième étape se décompose ainsi :

sélection des différents noms dont l'adjectif de référence est épithète dans un corpus de textes

constitution d'une matrice dont les lignes correspondent aux noms évoqués ci-dessus, les colonnes aux synonymes de l'adjectif de référence et les valeurs prises en entrées au nombre de cooccurrences (c'est-à-dire d'apparition conjointe) du couple (mot ; adjectif synonyme).

à partir d'une hypothèse d'équiprobabilité des distributions, calcul de coefficients théoriques selon le même procédé que dans le test du χ^2.

application d'une fonction linéaire par morceau (nulle, croissante, puis constante) au rapport $\frac{m_{ij}}{n_{ij}}$,

où m_{ij} est la valeur théorique et n_{ij} la valeur réelle, pour tous les couples (i,j) correspondant aux couples (mot ; synonyme). La valeur prise par la fonction, comprise entre 0 et 1, est qualifiée de degré d'affinité par Victorri.

Cette deuxième étape m'a paru particulièrement pertinente dans le cadre de mes travaux. Elle présente en effet plusieurs intérêts : elle dérive d'un test statistique de référence ; elle part de cooccurrences observées en cotexte, dans un corpus de textes ; elle affecte à tout couple de la matrice un coefficient d'affinité, contrairement au test du χ^2 qui additionne tous les écarts entre valeurs réelle et théorique et ne retourne qu'un coefficient global pour l'ensemble de la matrice ; le coefficient d'affinité repose sur la valeur relative du coefficient réel au théorique et s'affranchit de la valeur absolue (bien que la fonction choisie, linéaire par morceau dans ce cas, ne soit pas un élément indiscutable du modèle).

Néanmoins, soulignons quelques points importants : le cadre d'application du modèle de Victorri est très différent de celui dans lequel j'évolue. En effet, il se situe au niveau des mots, c'est-à-dire au niveau lexical. La notion de cooccurrence correspond à la relation nom – adjectif épithète. Dans mes démarches, la cooccurrence ne reposera pas sur la syntaxe mais sur la présence au sein d'une même unité textuelle. Cependant, à cette différence près, le cadre dans lequel je me placerai sera similaire.

3.2.2.2) Le modèle LSA

Le modèle LSA (analyse sémantique latente), développé par [Landauer, Foltz & Laham, 1998], est une théorie et méthode d'extraction et représentation du sens des mots en contexte par des traitements

statistiques appliqués à de larges corpus de textes. L'idée qui le sous-tend est que les contraintes mutuelles exercées entre mots dans des cotextes suffisent à faire émerger le sens. Ses objectifs se situent à deux niveaux : d'une part, il cherche à établir une similarité entre mots, par exemple pour déterminer si un mot peut être substitué à un autre ; d'autre part, il constitue un modèle de la réflexion et des démarches de la pensée pour acquérir et utiliser la connaissance.

LSA ne se construit que par analyse de textes. Son point de départ est un grand corpus, de trois millions de mots environ. Dans ce corpus, les mots sont assimilables aux points d'un « espace sémantique » de grande dimension (entre 50 et 1500). Les phrases ou encore les paragraphes, c'est-à-dire les cotextes choisis, correspondent aux dimensions de cet espace sémantique. Ces cotextes constituent des expressions unitaires de sens, pour lesquelles l'ordre n'est pas pris en compte : ni l'ordre des mots au sein du cotexte, ni l'ordre des cotextes entre eux. Seule compte la présence d'un mot dans un cotexte. Elle se traduit mathématiquement par la fréquence. Celle-ci subit un prétraitement, qui s'appuie sur la distribution du mot dans les cotextes où il est utilisé, indépendamment de ses corrélations avec d'autres mots et est transformée en une mesure de l'information qu'elle apporte. L'étape suivante, mécanisme clé de LSA, repose sur la réduction de la dimension des relations entre mots et cotextes. Enfin, une mesure de similarité entre deux mots est introduite.

Détaillons l'approche mathématique des étapes décrites ci-dessus :

Soit un corpus constitué de n mots $m_{i,1\le i\le n}$ et p cotextes $c_{j,1\le j\le p}$.

Soit f_{ij} la fréquence d'apparition du mot m_i dans le cotexte c_j.

Soit M la matrice des fréquences d'apparition des mots par cotexte :

$$M = \begin{array}{c} \\ m_i \end{array} \begin{array}{c} \begin{array}{c} c_j \end{array} \\ \begin{bmatrix} \vdots \\ \ldots \quad f_{ij} \\ \\ \end{bmatrix} \end{array}$$

Soit p_{ij} la probabilité d'apparition du mot i dans le cotexte j, P la matrice des p_{ij}.

Pour pondérer les coefficients en fonction de leur significativité, on applique la fonction $\pi : (f, p) \mapsto \begin{cases} \dfrac{\ln(f+1)}{-p\ln(p)} & si \quad p \neq 0 \\ 0 & si \quad p = 0 \end{cases}$ à tous les couples $\left(f_{ij}, p_{ij}\right)_{1\le i\le n, 1\le j\le p}$.

On note $\pi_{ij} = \pi\left(f_{ij}, p_{ij}\right)$ et Π la matrice des π_{ij}.

La réduction du nombre de degrés de liberté s'effectue par décomposition en valeurs singulières de la matrice Π : $\Pi = UDV^T$, $D = diag\left(\lambda_1, \lambda_2, ..., \lambda_q, 0, ..., 0\right), \lambda_1 \ge \lambda_2 \ge ... \ge \lambda_q$ diagonale de dimension (p,p) et de rang q, U orthogonale de dimension (n,p) et V orthogonale de dimension (p,p).

On souhaite se ramener à un sous-espace de dimension k, $k \le q$.

Soit $D_2 = diag\left(\lambda_1, \lambda_2, ..., \lambda_k, 0, ...0\right)$ et $\Pi_2 = UD_2V^T$.

L'opération effectuée correspond en fait à une projection sur k directions principales. Le mécanisme de pensée correspondant à cette transformation serait le suivant : l'esprit humain ne peut gérer la trop grande multiplicité de sens. C'est pourquoi il se ramène à des 'grandes lignes', c'est-à-dire des directions principales de sens. Les coefficients de Π_2 sont donc des coefficients de significativité des mots après transformation par l'esprit humain des textes.

La mesure de similarité est calculée à partir du cosinus des angles entre vecteurs-lignes de la matrice Π_2.

La démarche de LSA est intéressante à plusieurs points de vue. Tout d'abord, elle s'appuie sur des corpus, donc du texte vivant et non une ressource dictionnairique ou encyclopédique figée. Par ailleurs, elle s'appuie sur des méthodes et théories mathématiques solides : elle dérive de l'ACP, très utilisée en

analyse des données, et, avec l'introduction de l'entropie ($-p\ln(p)$), elle s'appuie sur la théorie de l'information.

Toutefois, l'analyse sémantique latente n'est pas parfaite et quelques points méritent attention. Elle se situe, comme les autres modèles mentionnés, au niveau lexical. Le modèle de réflexion humaine semble cohérent et rejoint une branche de la linguistique, la linguistique de l'interaction, pour laquelle l'apprentissage de la langue s'effectuerait à travers les échanges, les dialogues, la confrontation à des situations. Mais la théorie pose des problèmes de validation : elle n'a pour l'instant donné lieu qu'à des expériences à petite échelle, très ciblées et donc éloignées des interactions réelles.

3.2.2.3) Une tentative d'exploitation de plusieurs modèles : travaux de Mauceri

[Mauceri, 2007a et 2007b] utilise un modèle qui intègre différentes méthodes et théories précitées. Il se place dans une optique d'indexation de textes recourant à des rapprochements entre différents textes et s'appuie sur le repérage d'isotopies. Il se place ainsi dans le cadre de la sémantique textuelle. Pour quantifier son approche, il introduit une métrique dont l'angle d'observation est celui des cooccurrences, c'est-à-dire de l'apparition commune de deux traits sémantiques.

Le modèle qu'il bâtit puise ses sources dans une approche vectorielle, le modèle de Salton. Celui-ci décrit les cotextes comme des vecteurs dans l'espace des mots. La représentation matricielle des vecteurs est, de manière analogue au modèle décrit au paragraphe précédent une matrice dont les lignes sont les mots, les colonnes sont les cotextes et l'entrée en position (i,j) de la matrice le nombre d'occurrences du mot i dans le cotexte j. Mauceri souligne ensuite les failles de ce modèle et propose différentes manières d'y remédier. Il choisit en particulier de ne conserver que les mots sémantiquement pleins[2]. Une des méthodes qu'il propose ensuite est la transformation tf-idf appliquée à la matrice d'occurrences. Il complète cette transformation par des opérations similaires à celles de LSA. Il obtient ainsi une matrice dont les coefficients reflètent non plus les occurrences, mais des occurrences pondérées par leur significativité, avec prise en compte de la loi de Zipf et ses répercussions, ainsi que du modèle cognitif de LSA.

Il se ramène ensuite à un espace indépendant du découpage initial en textes ou cotextes. Pour ce faire, il génère une matrice de cooccurrences en multipliant la matrice précédente par sa transposée. Les nouveaux coefficients ne conservent certes qu'une partie de l'information liée au découpage en cotextes mais permettent de s'affranchir du texte pour se placer dans un espace ne dépendant plus que des interactions entre mots.

Enfin, Mauceri pointe du doigt le problème de la significativité statistique limitée des coefficients et propose une méthode de filtrage. Celle-ci élimine les coefficients statistiquement non significatifs et réajuste les autres coefficients selon qu'ils sont plus ou moins significatifs statistiquement. Pour cela, il s'appuie sur le test de Fisher.

Considérons le tableau de contingence suivant :

	A	$\overline{A}$	$m\arg e$
B	x	$b-x$	b
$\overline{B}$	$a-x$	$n-a-b+x$	$n-b$
$m\arg e$	a	$n-a$	n

La classe A (resp. B) est constituée de a (resp. b) individus, la population totale est de n individus. x individus appartiennent à A et B, a-x (resp. b-x) à A (resp. B) seulement, n-a-b+x à aucune des deux classes.

[2] Les mots pleins sémantiquement sont des mots évoluant extrêmement rapidement dans le temps. Ils se réfèrent à des situations, des événements, des objets ou des individus et portent donc une partie essentielle du sens des phrases et des textes. Ce sont souvent des noms, adjectifs, verbes, parfois aussi des adverbes. Ils s'opposent aux mots dits « mots grammaticaux » ou « mots outils », qui eux évoluent peu dans le temps. Dans les phrases, ces mots se situent autour des mots sémantiquement pleins, ils servent de lien entre eux. Il s'agit de prépositions comme « à » ou « de », des déterminants, des pronoms, etc.

Sous l'hypothèse H0 d'indépendance des lignes et des colonnes, la distribution suit une loi hypergéométrique :

$$f\left(x,a,b,n\right)=\frac{\binom{a}{x}\binom{n-a}{b-x}}{\binom{n}{b}}=\frac{a!b!(n-b)!(n-a)!}{n!x!(b-x)!(a-x)!(n-a-b+x)!}$$

, où $f\left(x,a,b,n\right)$ est

$$=\frac{\Gamma\left(a+1\right)\Gamma\left(b+1\right)\Gamma\left(n-b+1\right)\Gamma\left(n-a+1\right)}{\Gamma\left(n+1\right)\Gamma\left(x+1\right)\Gamma\left(b-x+1\right)\Gamma\left(a-x+1\right)\Gamma\left(n-a-b+x+1\right)}$$

la probabilité d'avoir x individus appartenant aux classes A et B de taille respectives a et b dans une population de n individus.

La probabilité d'avoir plus de ν individus appartenant à A et B est :

$$p\left(T\left(x,a,b,n\right),x\geq\nu\right)=\sum_{x\geq\nu}f\left(x,a,b,n\right)$$

Si l'on se replace dans le cadre linguistique, le tableau de contingence considéré est maintenant de la forme :

	$c_{\bullet 1}$	…	$c_{\bullet j}$	…	$c_{\bullet p}$
$c_{1\bullet}$			⋮		
⋮			⋮		
$c_{i\bullet}$	…		c_{ij}		
⋮					
$c_{n\bullet}$					

c_{ij} est le nombre de cooccurrences des mots i et j.

$c_{i.}$ (resp. $c_{.j}$) est le nombre de cooccurrences du mot i (resp. j) avec l'ensemble des autres mots

c est le nombre total de cooccurrences : $c=\sum_i c_{i\bullet}=\sum_j c_{\bullet j}$

L'hypothèse H0 est que tous les couples de mots sont indépendants. La probabilité d'observer c_{ij} cooccurrences entre les mots i et j sachant que le mot i cooccurre $c_{i.}$ fois et le mot j $c_{.j}$ fois est de :

$$f\left(c_{ij},c_{i\bullet},c_{\bullet j},c\right)=\frac{\Gamma\left(c_{i\bullet}+1\right)\Gamma\left(c_{\bullet j}+1\right)\Gamma\left(c-c_{\bullet j}+1\right)\Gamma\left(c-c_{i\bullet}+1\right)}{\Gamma\left(c+1\right)\Gamma\left(c_{ij}+1\right)\Gamma\left(c_{\bullet j}-c_{ij}+1\right)\Gamma\left(c_{i\bullet}-c_{ij}+1\right)\Gamma\left(c-c_{i\bullet}-c_{\bullet j}+c_{ij}+1\right)}$$

La probabilité d'avoir plus de c_{ij} cooccurrences entre les mots i et j est :

$$p_{ij}=p\left(T\left(x,c_{i\bullet},c_{\bullet j},c\right),x\geq\nu\right)=\sum_{x\geq\nu}f\left(x,c_{i\bullet},c_{\bullet j},c\right)$$

Une cooccurrence sera considérée comme non significative si $p_{ij}\geq\alpha$, où α est le seuil de cooccurrences. La matrice des coefficients filtrées aura pour valeur $\frac{\alpha-pij}{\alpha}\times c_{ij}$ en position (i,j) si $p_{ij}\leq\alpha$ et 0 sinon.

La méthode de Mauceri m'a intéressée à plusieurs points de vue et je m'en suis inspirée dans mes propres démarches. Tout d'abord, elle traite les problèmes suivants : celui des termes trop fréquents et peu significatifs (méthode tf-idf) ; celui de la significativité statistique. Elle applique LSA et se place ainsi dans la mouvance du modèle cognitif qu'il propose. Elle opte pour une approche en cooccurrences plutôt qu'en occurrences, choix qui donne une marge de liberté par rapport au support de cotextes choisis et à l'extension à d'autres cotextes.

Notons cependant que, si chaque transformation considérée indépendamment des autres paraît pertinente et interprétable sur le plan linguistique, l'enchaînement des transformations a des effets plus difficiles à se représenter au niveau linguistique.

3.2.3) Autres perspectives

Il existe de nombreux modèles mathématiques qui ont été mis en œuvre pour de la linguistique ou dont découlent certaines transformations précédemment citées. Parmi elles, mentionnons l'ACP, intéressante pour son approche multidimensionnelle ; les chaînes de Markov, dont la dynamique ouvre des perspectives intéressantes (mais pose aussi le problème de la divergence) ; les réseaux de neurones ; des modèles de graphes entre synonymes, mots d'articles ou définitions de dictionnaire, … Je n'ai pas approfondi ces pistes à fort potentiel pour deux raisons : un temps trop court et un risque d'éparpillement.

Retenons de tous les modèles et transformations décrits que chacun présente des atouts séduisants, mais que, dans mes choix, j'ai donné la préférence à des modèles plutôt récents et qui m'ont paru le plus à même de répondre au cadre de mon étude. Dans tous les cas, il est nécessaire de faire la part des choses : si la théorie semble cohérente et donne des résultats probants dans un champ bien déterminé, mon cadre d'application (niveau infra-lexical et recours aux traits sémantiques) est différent, vierge d'expérimentation et exige certainement des adaptations que seule l'expérience pourra mettre en lumière.

3.3) Ressources informatisées et outils de traitement

La réflexion théorique et l'élaboration de modèles sont des démarches riches et constructives. Cependant, pour valider le théorique et orienter les pistes de réflexion, il est nécessaire de se confronter à la pratique. Or l'expérimentation requiert des ressources et des moyens techniques. L'ATILF proposait différents outils informatisés et différentes ressources informatisées susceptibles d'être exploités. Je présente ci-dessous des outils intéressants par rapport à mon sujet, j'analyse leur pertinence et j'explique pourquoi je les ai retenus ou non pour les expériences de ces quatre mois de stage.

3.3.1) Première ressource informatisée : un dictionnaire, le TLFi

Le TLFi contient de la matière pour générer des données et présente une structure favorable au développement d'outils d'exploitation. Il constitue donc un support fondamental sur le plan pratique, que j'ai exploité dans mes expériences.

Tout d'abord, il peut être considéré comme un réservoir de traits sémantiques, autrement dit, dans ce projet, le sémème de tout mot est assimilé à l'ensemble des mots sémantiquement pleins de sa définition (noms, verbes, adjectifs, adverbes). Cette hypothèse repose sur les arguments suivants. Un mot peut être considéré comme un ensemble de traits sémantiques. Les termes sémantiquement pleins de sa définition servent à faire émerger le sens de ce mot, on peut donc légitimement supposer qu'ils appartiennent à son sémème. Par ailleurs, la définition doit permettre d'appréhender le sens d'un mot inconnu quel que soit son contexte d'apparition. Le sémème est donc inclus dans les unités de sens véhiculées par les termes définitoires. Certes, le jeu sur le double niveau, lexical et infra-lexical, pose le problème des imbrications multiples : un terme de la définition peut être vu comme trait sémantique mais également comme mot, auquel cas il est lui-même composé d'un ensemble de traits sémantiques qui, eux-mêmes pris comme mots, sont constitués de traits sémantiques, etc. Nous partirons de l'hypothèse que les termes de la définition forment le sémème en première approximation.

Par ailleurs, le TLFi est un outil riche, relativement fiable et structuré. La richesse apparaît à travers la grande diversité lexicale (100000 mots, 270000 définitions) et l'objectif fixé d'exhaustivité sur les mots du XIXe et XXe siècle. Soulignons des lacunes sur le vocabulaire de la fin du XXe siècle. Sur ce point, des modifications sont actuellement en cours avec la réalisation du supplément du TLF mais ne

sont pas encore intégrées. A la diversité lexicale s'ajoute une richesse du contenu des entrées. Celles-ci comportent la définition même, mais également d'autres rubriques : exemples, titres, dates et auteurs d'exemples, constructions, syntagmes, domaines techniques, synonymes et antonymes, sources. Dans les expériences menées, seules les définitions ont servi à constituer le sémème affecté à un mot, mais l'existence des autres rubriques ouvre des perspectives d'enrichissement de ce sémème.

D'autre part, concernant la fiabilité et la structuration, le TLFi a été rédigé par des lexicographes pendant trente ans. Il est donc le fruit du travail approfondi de personnes qualifiées. La question de la compétence des rédacteurs n'est pas problématique comme dans le cas des wiki (site web enrichi et modifié par des utilisateurs).

Enfin, les informations du TLFi peuvent être facilement récupérées et exploitées. En effet, une version simplifiée du TLFi, appelée SEMEME, est issue de l'exploitation du codage XML du TLFi ainsi que de l'étiquetage grammatical des mots apparaissant dans les définitions. L'encodage XML de SEMEME permet d'accéder aisément au contenu comme à la structure. Cette ressource comporte notamment l'ensemble des mots sémantiquement pleins (noms, verbes, adjectifs, adverbes) composant une définition. Soulignons néanmoins que toutes les fonctionnalités ne sont pas contenues dans cette version XML. En particulier, la fonctionnalité de mots apparentés n'est pas conservée. Cette fonctionnalité permet de retrouver les mots susceptibles de correspondre à une entrée non identifiée, par exemple le singulier d'un pluriel (le mot apparenté à « désastres » est « désastre »), le remplaçant potentiel d'un mot mal orthographié (« hagard », « agar » ou « agare » pour l'entrée « agard »).

3.3.2) Bases textuelles

Les bases textuelles sont des supports essentiels car ce sont elles qui fournissent la matière pour constituer et procurent des cotextes d'un mot donné. Les bases textuelles présentées ci-dessous sont diverses, aussi bien en contenu qu'en structure informatique. Nous aborderons d'abord Frantext, base de textes littéraires, puis la base journalistique de l'Est Républicain, ensuite Wikisource, base de textes libres de droit disponible en ligne et enfin un outil capable de générer des corpus à partir du web : Pompadoc.

3.3.2.1) Frantext, une base de textes littéraires

L'interface web de Frantext (voir 2.é) permet de sélectionner des textes en fonction de ses besoins et ainsi de constituer des corpus. Les textes sont accessibles par auteur, titre, genre et dates. Une recherche peut être effectuée dans les textes par mots, mais aussi par critères plus sophistiqués : lemmes, expressions, liste de mots, entités catégorisées, séquence de mots, mots séparés par un certain intervalle,… Cette recherche complexe permet la constitution relativement fine de corpus, du moins si celle-ci s'élabore autour d'un ou plusieurs mots ou expressions.

Frantext présente donc une indéniable richesse et un mode de génération de corpus intéressant. Son contenu est littéraire, ce qui garantit une certaine qualité de construction des textes. La probabilité d'avoir dans les écrits des relations sémantiques pensées et non fruits d'une maladresse est plus grande qu'en langue parlée ou des textes récupérés sur des blogs. En revanche, Frantext présente des inconvénients assez sérieux. La mise en forme des textes de Frantext respecte la disposition d'origine des textes dans les ouvrages. Ainsi, les textes saisis sont balisés en fins de ligne : celles-ci sont les mêmes que celles des ouvrages d'origine. Au contraire, phrases et paragraphes ne sont pas balisés. Or, par rapport à mon axe d'approche, la structuration du texte en unités sémantiques et non lexicales est fondamentale. Autre problème, en partie lié au précédent : une recherche centrée sur un mot permet de visualiser un cotexte de ce mot qui n'est pas nécessairement de taille appropriée. Ce cotexte a en effet une taille indépendante de la structuration en paragraphe ou autre unité sémantique à laquelle il appartient. De plus, sur le plan pratique, la récupération du cotexte n'est pas automatique mais manuelle (copie du cotexte apparaissant à l'écran dans un nouveau document texte). Ajoutons enfin que les genres de Frantext ne sont pas clairement définis, problème actuellement en cours de traitement au sein de l'ATILF. Cette faiblesse influe sur l'homogénéité et la qualité du corpus constitué.

Frantext a dans un premier temps été le candidat principal à la constitution de corpus pour mes expériences, mais les raisons techniques mentionnées ci-dessus l'ont relégué au second plan.

3.3.2.2) L'Est Républicain, corpus de textes journalistiques

L'ATILF dispose d'un corpus journalistique constitué d'articles de l'Est Républicain. Ce corpus comporte l'ensemble des articles parus en 1999, 2002 et 2003. Il est disponible au format TEI (la TEI, Text Encoding Initiative, est une norme de balisage, de notation et d'échange de corpus). La structure interne comporte notamment un balisage en articles décomposés en une accroche, un titre et le corps de l'article ainsi qu'un balisage en paragraphes, non systématique mais assez fréquent.

Un des avantages de ce corpus est qu'il est ancré dans l'actualité et correspond à une pratique sociale bien déterminée à savoir celle du discours journalistique. De plus, son contenu est très différent des textes littéraires de Frantext. Il permet donc une approche complémentaire particulièrement utile. En effet, des résultats concluants d'un modèle mathématique sur un corpus de textes ne garantissent pas l'universalité de ce modèle. La comparaison des résultats obtenus dans des corpus de nature différente permet de faire émerger des failles du modèle ou d'entériner sa robustesse.

Signalons quelques points critiques. Les articles disponibles à l'ATILF ne sont pas les versions définitives et comportent parfois des commentaires des rédacteurs à certains emplacements. De plus, les informations ne sont pas toujours dans la bonne catégorie. Par ailleurs, il est actuellement impossible d'effectuer une sélection d'articles par mot-clé.

Malgré l'approche intéressante de la langue qu'il offre, l'Est Républicain n'a pas servi de support dans la phase expérimentale. Outre les questions techniques mentionnées ci-dessus, il présente un autre inconvénient, lié au choix du TLFi comme ressource de référence. En effet, le vocabulaire de l'Est Républicain est celui des années 2000. Or le TLFi n'intègre pas le vocabulaire récent et s'arrête vers les années 90. L'introduction au Supplément du TLFi doit remédier à ce problème mais n'est pas encore effective. Il paraissait donc plus judicieux de différer l'exploitation de l'Est Républicain à la mise en place du Supplément.

3.3.2.3) Wikisource, des contes parmi un vaste panel de textes

Wikisource (http://fr.wikisource.org/wiki/Accueil) est une bibliothèque libre en ligne. Elle est constituée d'environ 10000 textes de 1700 auteurs. Ces textes sont sous licence libre ou passés dans le domaine public et se répartissent en différentes catégories : littérature, sciences humaines, exactes et sciences de la nature, religion, arts. Les textes sont accessibles par genre, époque, auteurs, livres, courants, thèmes ou encore mots-clés.

Les textes mis à disposition par Wikisource présentent plusieurs intérêts : ils offrent de la variété, sont libres de droits et déjà mis en forme.

Ils présentent toutefois quelques inconvénients. Ainsi, les outils de sélection des textes sont moins développés que ceux de Frantext. En outre, les textes sont récupérables par des moyens manuels (copier-coller) mais ne sont pas disponibles en XML. Cette situation est gérable pour un corpus de textes de taille limitée, mais est plus problématique pour de grands corpus. Enfin, Wikisource est un wiki, donc modifiable par tout utilisateur, c'est-à-dire très évolutive. On peut certes supposer que sa stabilité est plus importante que celle des pages web accessibles par les moteurs de recherche. Cependant, l'évolution des textes (ajout, suppression, modification de la mise en forme comme par exemple le découpage en paragraphes) n'est pas contrôlable. La reproductibilité d'une expérience peut donc de ce fait être mise à mal.

En raison de ses atouts forts et malgré les bémols signalés, j'ai opté pour des textes, plus précisément des contes de Wikisource dans mes applications. En effet, les contes m'ont paru un genre particulièrement favorable à l'analyse (Les contes4.2).

3.3.2.4) Corpus constitué à partir du web par le biais de l'outil Pompadoc

La Pompadoc, développée au sein de l'ATILF, prototypée par Jérémie Ceintrey et Yorick Petey et maintenue par Sandrine Ollinger, est un outil d'aspiration et de stockage de pages web à partir de moteurs de recherche (actuellement Yahoo et Google).

Elle sélectionne les pages web à partir de mots-clés, après diverses spécifications : langue des sites, nombre de pages à aspirer, taille minimale ou maximale en mots des pages, éliminations des pages en double et éventuellement nom du domaine au sein duquel effectuer la recherche. Sur Google, il est également possible de préciser l'emplacement où le mot doit être localisé : adresse URL, titre ou texte. Une fois les pages sélectionnées, elles sont récupérées au format HTML et converties au format XHTML. L'opération suivante vise à conserver et structurer les informations ad hoc, puis à récupérer ces informations sous format XML/TEI. Elle s'appuie sur l'utilisation de feuilles de style. Par exemple, pour des sites de presse régionale et nationale, l'article principal, le titre et l'auteur sont récupérés et les commentaires d'utilisateur, les images ou encore la publicité éliminés.

Un tel outil apparaît comme extrêmement précieux dans une optique de constitution de corpus. Son champ d'investigation est vaste, il peut collecter des pages web issues de la presse aussi bien que de blogs ou de sites officiels.

Toutefois, une critique essentielle peut être formulée à l'encontre de Pompadoc. Elle concerne la fluctuation des informations disponibles sur la Toile : le web est en évolution permanente et les pages aspirées aussi bien que leur contenu peuvent changer d'un jour à l'autre. Ceci soulève le problème de la reproductibilité des expériences, critère fondamental en sciences.

3.3.3) Deux outils récemment développés pour la sémantique textuelle : regroupements morphologiques et Sémy

3.3.3.1) Regroupements morphologiques

François Rastier propose d'analyser le sens suivant trois paliers : microsémantique (mot), mésosémantique (du syntagme à la période, unité regroupant plusieurs syntagmes et inférieure au texte) et macrosémantique (texte).

[Ramdani, 2007] s'attache à regrouper les traits sémantiques à partir d'analyses microsémantiques. Elle constitue des familles de traits sémantiques à partir de leur structure morphologique interne. Elle détermine des critères théoriques de regroupements et met en place un outil informatique les réalisant.

Les regroupements se fondent sur différentes méthodes. La méthode fondamentale repose sur l'analogie graphique de mots, c'est-à-dire sur des rapprochements à partir de la similarité de la séquence de lettres composant le mot. Ainsi, banane et bananier présentent une analogie graphique, de même que retranscrire et transcription ou angle et anglais. Pour éviter des regroupements malencontreux, comme le dernier exemple mentionné, elle recourt au TLFi. Son hypothèse est la suivante : un mot graphiquement proche d'un autre mot et ayant un lien sémantique avec celui-ci aura tendance à apparaître dans sa définition. Elle ajoute d'autres méthodes à celle-ci, afin d'augmenter le rappel (rapport du nombre d'éléments pertinents sélectionnés sur le nombre total d'éléments pertinents) : elle utilise le lexique morphologique Verbaction qui à un verbe associe les noms d'action correspondants et exploite les résultats de l'analyseur morphologique DériF qui travaille sur les suffixes (-tion, -able par exemple), préfixes (re-, in-,…) et effectue des conversions adjectif – verbe.

Les regroupements morphologiques effectués permettent de passer d'un peu plus de 40000 sèmes à plus de 7000 familles et près de 22000 sèmes non regroupés. Voici un exemple de famille obtenue :

Famille du sème /bicyclette/ *:* cyclisme,*NOM* cyclotourisme,*NOM* bicyclette,*NOM* cyclotourisme,*NOM* cyclable,*ADJ* cycliste,*ADJ* bicycliste,*NOM* bicycle,*NOM* cycliste,*NOM*

Ces regroupements présentent un intérêt majeur : ils réduisent le nombre d'éléments distincts, ce qui permet de lutter contre une forme de dispersion et évite d'obtenir des matrices encore plus creuses que celles obtenues lors des expériences menées au cours de ce stage.

Cependant, les regroupements ne sont pas tous satisfaisants et doivent être manipulés avec précaution. En effet, si la plupart des regroupements paraissent appropriés, d'autres sont trop larges ou non pertinents, générant des familles dont le cœur sémantique est parfois difficile à dégager. Citons par exemple la famille de /forme/ qui comporte 182 items, parmi lesquels informatique, effort, réforme, formule ou encore formaliste. D'autres regroupements, plus petits et moins hétérogènes, n'en

sont pas moins problématiques. L'homonymie[3] et la polysémie en sont à l'origine, comme dans le cas de /chanter/, dont le regroupement comportera aussi bien chantage que chanteur, ou action regroupant actionnaire et activisme. Ces problèmes ont été repérés, leur cause analysée et une description théorique de regroupements plus fins existe, du moins partiellement, mais elle n'a pas été mise en œuvre informatiquement.

Mon équipe de travail, tout en ayant conscience des limites, s'est accordée pour conserver les regroupements à la fois pour les avantages qu'ils présentaient et pour des raisons techniques d'espace mémoire insuffisant. L'utilisation des regroupements actuels est une solution temporaire qui exige, à terme, un travail d'affinage.

3.3.3.2) Sémy

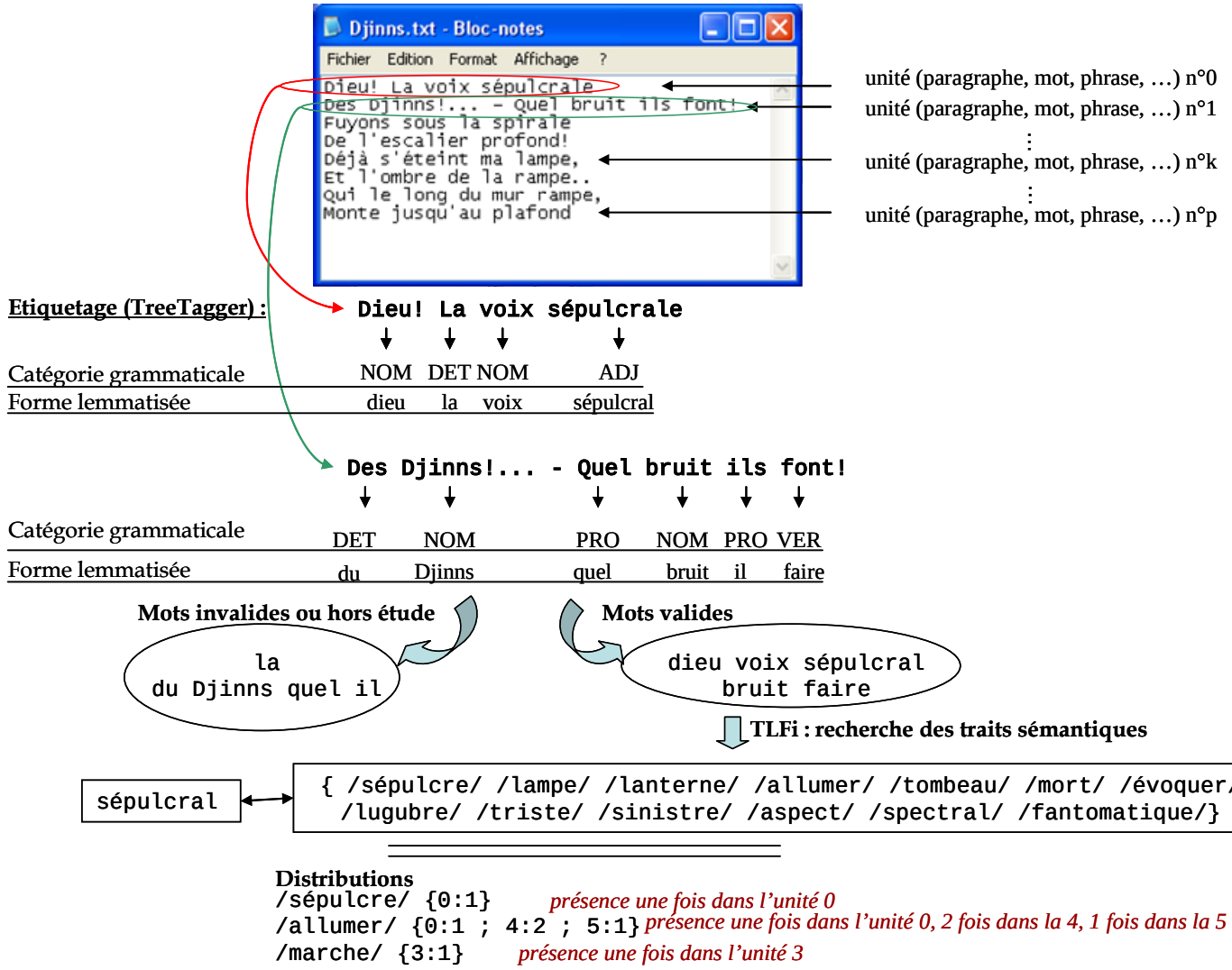
Sémy est une plateforme d'annotation en traits sémantiques. Il s'agit d'un programme informatique écrit en Python réalisé par [Grzesitchak, 2007] dans le cadre de la sémantique textuelle. Il associe à des unités de textes (mots, phrases, paragraphes) les traits sémantiques correspondants et leur nombre d'occurrences.

Voici une brève description de son mode de fonctionnement. Le point de départ est un texte découpé en unités (paragraphe, phrase, fenêtre de mots, …). On souhaite connaître les traits sémantiques présents dans le texte et leur distribution, c'est-à-dire leur nombre d'occurrences par unité de découpage. Sémy prend en entrée un fichier texte où est écrit le texte de référence. Ce fichier comporte une unité (paragraphe, …) par ligne. Pour chaque ligne du texte, Sémy détermine la catégorie grammaticale et la forme lemmatisée[4] des mots de cette ligne grâce à TreeTagger, système automatique d'étiquetage grammatical et de lemmatisation. Il trie ensuite les lemmes : seuls sont conservés les noms, verbes, adjectifs et adverbes : les autres catégories sont écartées pour le moment. Les lemmes sont ensuite recherchés dans SEMEME, la version simplifiée du TLFi. Certains, non identifiés, soit à cause d'un mauvais étiquetage de TreeTagger, soit parce qu'ils sont absents de SEMEME, sont considérés comme invalides et éliminés. A chaque lemme restant sont affectés les mots sémantiquement pleins de la ou des définition(s) de SEMEME correspondante(s), considérés comme les traits sémantiques. Sémy se base en effet sur l'hypothèse que les traits sémantiquement pleins de la définition constituent le sémème de l'entrée du dictionnaire choisie. Il peut regrouper ensuite les traits sémantiques en familles morphologiques (voir paragraphe précédent), puis établit la distribution des traits sémantiques.

Schéma illustrant le fonctionnement de Sémy :

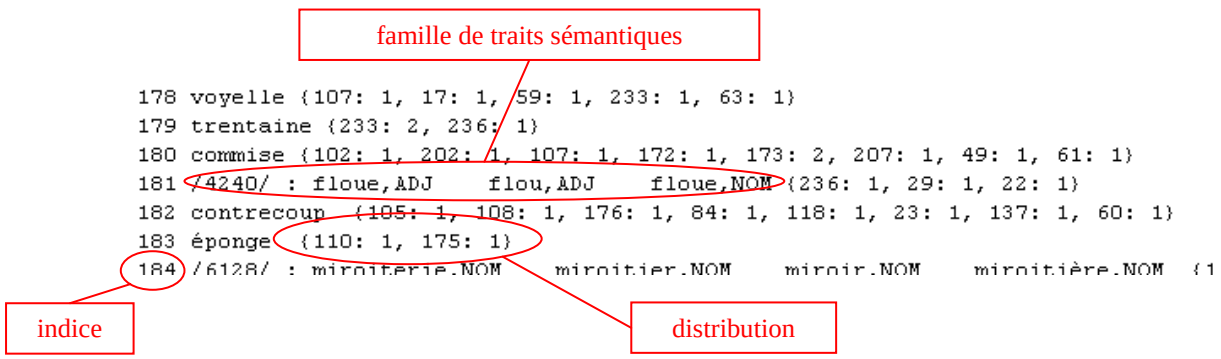
[3] Relation entre plusieurs formes linguistiques ayant le même signifiant graphique et/ ou phonique et des signifiés totalement différents. Ex : avocat (magistrat et fruit) ; mère et mer.

[4] La lemmatisation d'une forme est la mise sous forme conventionnelle de celle-ci, son *lemme*, et correspond à son entrée dans un lexique (par exemple, la forme lemmatisée d'un verbe conjugué sera son infinitif ; le lemme d'un adjectif au féminin pluriel, comme *sucrées*, sera le masculin singulier, *sucré*).



[5]

En sortie, Sémy retourne plusieurs fichiers : des fichiers annexes sur l'étiquetage de TreeTagger et les termes éliminés (la liste des mots hors étude, des mots invalides, des mots étiquetés grammaticalement et lemmatisés ou encore des mots conservés) et des fichiers centraux, avec en particulier un fichier aux formats csv et html qui indique les familles de traits sémantiques apparus dans le texte, un indice qui leur est affecté arbitrairement et la distribution de la famille de traits par unité.



[5] Schéma fictif dont la vocation est purement pédagogique.

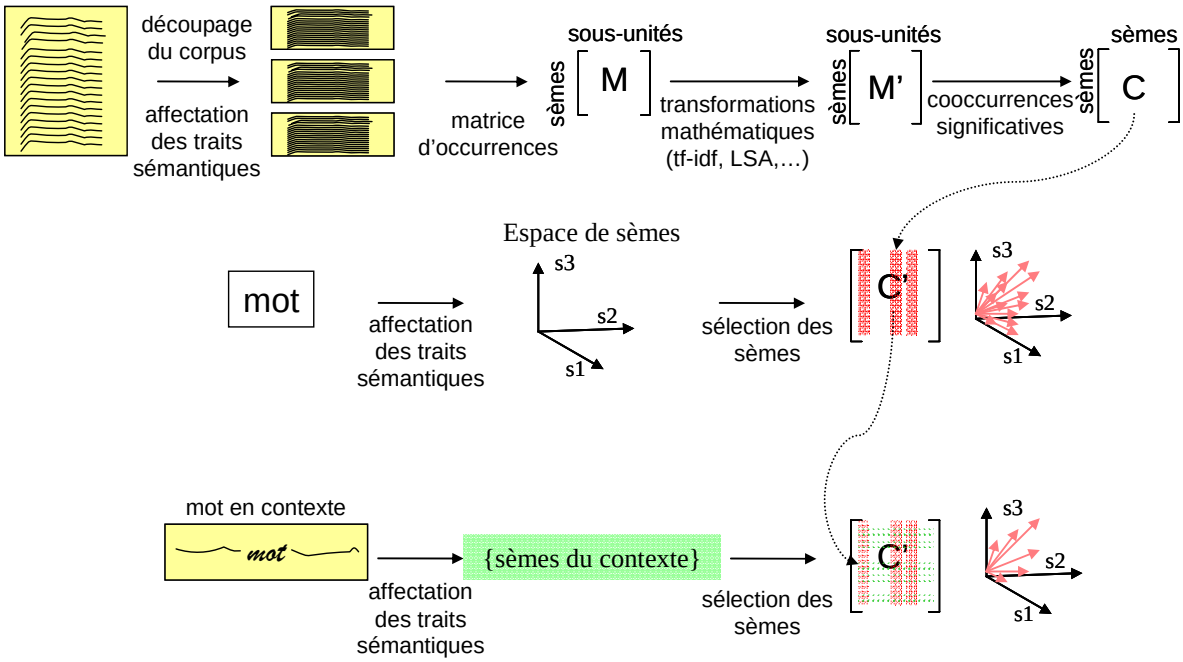
Sémy possède un certain nombre d'options et d'extensions qui n'ont pas été exploitées mais méritent d'être mentionnées. Une première option consiste à choisir entre un décompte simple (présence / absence : 1 ou 0) ou multiple des occurrences par définition. Ainsi, si le trait /allumer/ apparaît deux fois dans la définition de lampe et une fois dans la définition d'éteindre, le nombre d'occurrences sera de deux en décompte simple (1+1) et de trois (2 +1) en décompte multiple dans l'unité *Déjà s'éteint ma lampe*. Autre option proposée par Sémy : effectuer ou non les regroupements morphologiques de [Ramdani, 2007]. Le programme a toujours été utilisé en mode regroupement dans les expériences menées pour les raisons invoquées au paragraphe précédent. Par ailleurs, Sémy peut prendre en compte un double découpage, par exemple découpage d'un corpus en textes eux-mêmes découpés en paragraphes. Pour signaler le double découpage, il faut constituer autant de fichiers .txt qu'il y a de textes et organiser les fichiers .txt comme décrit précédemment, avec un paragraphe par ligne. Sémy retourne alors un fichier comportant la distribution des traits sémantiques par textes et interne à chaque texte. Supposons que le fichier de sortie indique une distribution de la forme {0 :{0 :1 ; 2 :1} ; 1 :{0 :3} ; 2 :{1 :4}} pour le trait sémantique /cornaline/: cette notation signifie que le trait est présent deux fois dans le texte n° 0 (une fois au paragraphe 0 et une fois au paragraphe n°2), trois fois dans le texte n° 1 (dans le paragraphe 0), quatre fois dans le texte n° 2 (au paragraphe n°1). Cette fonctionnalité n'a pas été exploitée mais est riche de perspectives si on se place dans une optique de découpages imbriqués. Par ailleurs, Sémy ne se contente pas de déterminer des distributions, il établit également certains calculs statistiques, à savoir moyenne et écart-type calculés à partir des distributions.

Signalons enfin une caractéristique actuelle de Sémy : pour l'instant, le sémème affecté à partir du TLFi est non réflexif, autrement dit le mot qui sert d'entrée n'est pas intégré à son propre sémème.

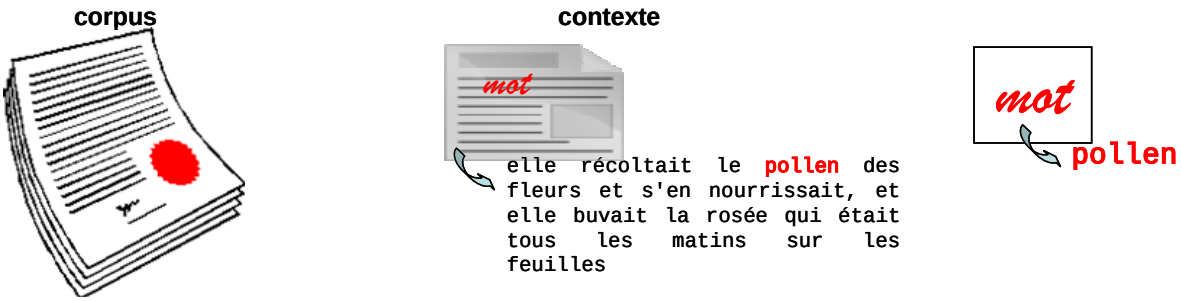
IV) Modèle optimal

4 .1) Démarche globale

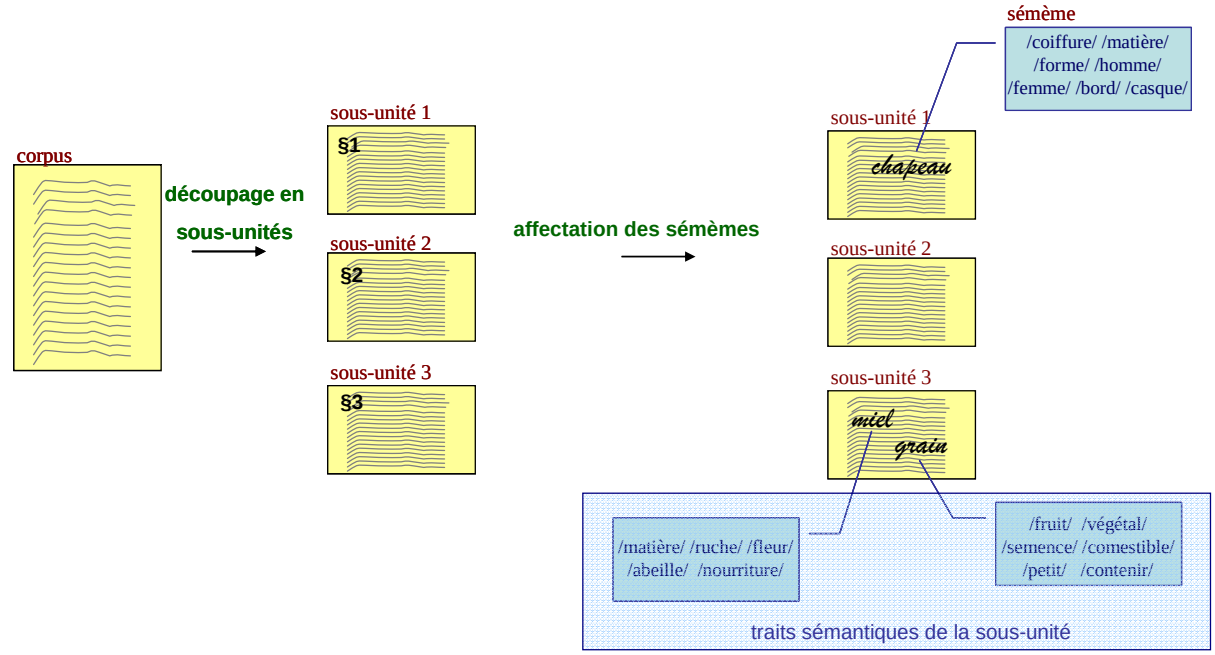
Notre objectif est dans un premier temps de parvenir à une image mathématique globale, obtenue à partir d'un corpus de textes. Cette image doit refléter les affinités entre traits sémantiques. Il s'agit dans un deuxième temps d'extraire de cette représentation mathématique globale une image locale, c'est-à-dire centrée sur un mot de référence et le cotexte proche de celui-ci. Nous qualifierons cette démarche de modélisation et prions le lecteur de ne pas voir derrière le terme de modèle des prétentions plus ambitieuses.



L'étape préalable est la sélection d'unités textuelles appropriées (partie 4.2)). Elle correspond au choix d'un corpus de textes, du mot dont on souhaite étudier les variations sémantiques et du cotexte d'apparition de ce mot. Le choix du corpus, du mot de référence et du cotexte est conditionné par les expériences que l'on souhaite effectuer.

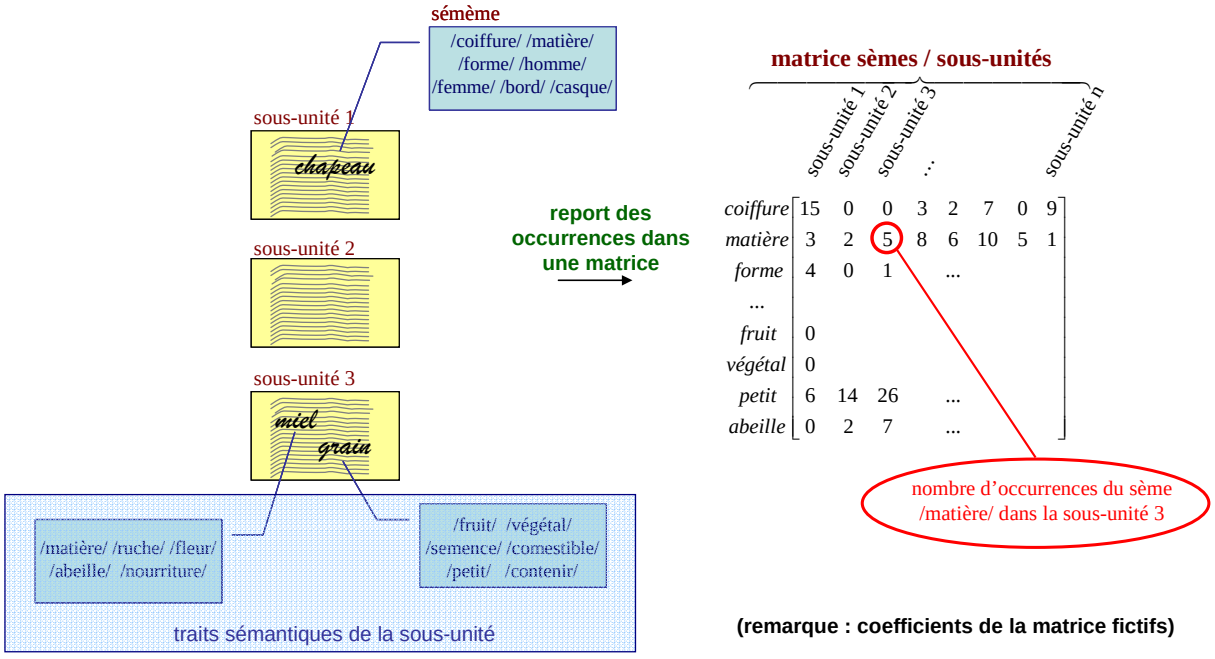


Les unités textuelles sont ensuite mises à un format standard (partie 4.3) : le corpus doit être structuré, c'est-à-dire découpé en sous-unités. Chaque unité et sous-unité textuelle est ensuite transposée du plan lexical au plan infra-lexical : son sémème lui est affecté.



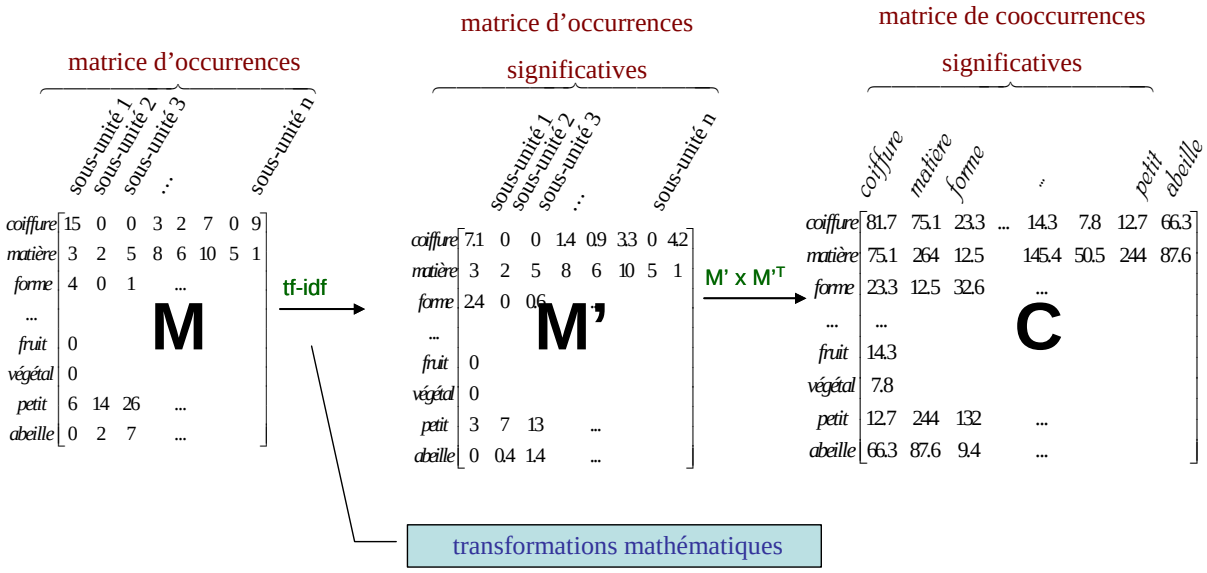
Au stade suivant (partie 4.4), l'image du corpus obtenue par passage du niveau lexical au niveau infra-lexical est convertie en une représentation mathématique. Le passage du qualitatif au quantitatif repose sur le dénombrement d'occurrences, c'est-à-dire d'apparitions, des traits sémantiques. Les sous-unités du corpus constituent le support de distribution des occurrences : les apparitions d'un trait sémantique sont décomptées par sous-unités.

Ces décomptes permettent de générer une matrice d'occurrences. Les lignes de la matrice correspondent aux traits sémantiques, les colonnes aux sous-unités du corpus et les entrées de la matrice au nombre d'occurrence de chaque trait sémantique par sous-unité.



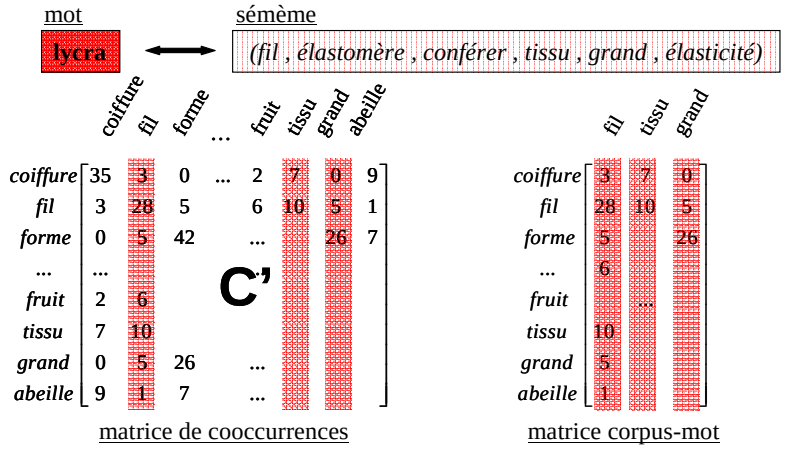
(remarque : coefficients de la matrice fictifs)

Par la suite, cette matrice subit des transformations mathématiques à double vocation : convertir le nombre d'occurrences en un coefficient de significativité ou de proximité sémantique ; se ramener à un espace ne dépendant que des traits sémantiques et de leurs relations réciproques (génération d'une matrice dite de cooccurrences).

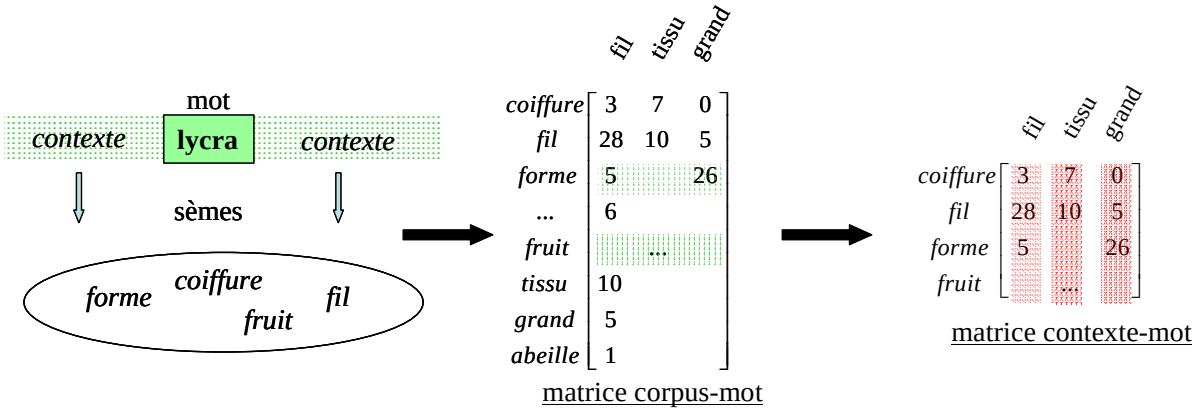


La matrice finale s'interprète comme une représentation sémantique du monde telle qu'elle apparaît à travers le corpus : elle doit refléter l'importance que nous accorderions à un trait sémantique a priori et les associations entre traits sémantiques que nous effectuerions hors de tout contexte. Cette représentation matricielle du monde est tributaire du corpus de départ : les traits sémantiques /avocat/ et /agricole/ auront un poids de cooccurrences différent si le corpus de départ est un ensemble de textes juridiques ou un ensemble de rapports d'activité en agro-alimentaire. A titre d'illustration, on peut aisément imaginer qu'un enfant élevé aux contes et aux dessins animés n'effectuera vraisemblablement pas les mêmes associations sémantiques qu'un étudiant chinois en ingénierie initié à la langue française par le vocabulaire de l'entreprise.

Ensuite, la représentation d'un mot donné est extraite de l'image sémantique du corpus que l'on peut voir comme une « toile sémantique » de ce corpus. Cette approche permet de prendre en compte la représentation globale, c'est-à-dire l'ensemble du corpus, dans la représentation locale d'un mot. Une approche complémentaire et mathématiquement équivalente est la suivante : le corpus est vu sous un angle d'observation particulier, à savoir à du point de vue du mot.



A cette étape succède le passage du global au local : de la représentation du mot sur l'ensemble du corpus n'est conservée que la projection sur le sous-espace correspondant au cotexte du mot, ou, dans l'optique de l'approche complémentaire, seule la partie représentative du cotexte est conservée dans l'espace du mot.



La réitération des étapes précédentes pour divers cotextes, mots, corpus ou transformations mathématiques fournit des images sémantiques différentes. Des traitements mathématiques complémentaires sont alors nécessaires pour accéder à une représentation synthétique des données ou visualiser celles-ci. Ces processus ouvrent la porte à l'analyse et la comparaison de résultats.

4.2) Choix des matériaux de base

Les éléments de départ sont les suivants : le corpus, le mot et son cotexte. La sélection est guidée par un certain nombre de critères. Cependant, le choix optimal est difficile à définir : nombre de critères se heurtent à des controverses, le passage du qualitatif au quantitatif n'est pas clairement défini et la linguistique, science des nuances, doit composer avec des mathématiques réclamant du général et le lissage des cas particuliers.

La constitution du corpus soulève des questions autour des points suivants : l'homogénéité, la taille et la structuration interne du corpus.

Un corpus doit-il être homogène ou hétérogène ? Les informations qu'il fournira serviront, rappelons-le, à créer la 'représentation du monde' véhiculée par la matrice finale du corpus. Cette représentation fait écho aux reliefs et connexions sémantiques sous-jacentes au cerveau humain. Un tel argument plaide en faveur de l'hétérogénéité. Cependant, l'hétérogénéité bute contre la contrainte des limites et peut prendre différentes formes : genre, époque, domaine ou encore thème.

Avons-nous les mêmes intuitions sémantiques lorsque nous abordons du théâtre, des essais ou de la poésie ? Non, car les discours (discours littéraire, scientifique, …) correspondent à des pratiques sociales. Les pratiques sociales correspondent elles-mêmes à des usages linguistiques. De plus, au sein d'un discours, les genres textuels répondent à des codes de rédaction et d'interprétation, contraintes intégrées à la fois par l'auteur et le lecteur. Le sens est donc en partie fonction du genre. Dans le cadre de cette étude, les genres suivants ont été présélectionnés : des contes (voir Wikisource), des romans et un corpus journalistique. Les contes sont susceptibles de donner une représentation du monde telle que les adultes la destinent aux enfants. Les notions abstraites y sont sous-représentées mais le caractère imagé, concret, facilite l'analyse et en fait un candidat légitime aux premières expériences. Le corpus littéraire offre une garantie de structure : les auteurs littéraires veillent généralement à respecter l'unité sémantique d'une phrase, d'un paragraphe ou encore d'un chapitre. Un bémol toutefois : les effets de style et le caractère extrêmement construit des ouvrages n'éclipsent-ils pas le rôle déterminant de l'usage, notion au cœur de ma thématique ? Reste à savoir comment définir les limites de l'usage. Quant au corpus journalistique, il constitue un panel certainement plus proche de l'usage, même si les champs abordés sont parfois très abstraits et moins propices à l'analyse, du moins pour les débuts. Soulignons enfin que cette présélection a aussi été guidée par les outils disponibles.

Comment choisir les limites temporelles ? En effet, le sens des mots évolue dans le temps : la situation du mari déçu de la championne sportive détrônée et du mari déçu dans les comédies de Molière sont loin d'être comparables. Un critère invite à choisir une époque relativement récente : les

définitions du TLFi constituent le réservoir de traits sémantiques. Or ces définitions ont été rédigées par des linguistes du XXe siècle pour leurs contemporains. De plus, le TLFi n'a pas vocation à être un dictionnaire historique. Toutefois, la notion de « relativement récent » reste à définir. Un texte du XVIIIe l'est-il ? Du XIXe ? La limite suivante a été retenue : les textes choisis sont postérieurs au XIXe siècle et antérieurs aux années 90, à partir desquelles nous savons que les informations à venir du Supplément au TLFi nous serons nécessaires..

Une fois le genre et l'époque sélectionnés, une certaine hétérogénéité est à privilégier. Il est judicieux de ne pas se centrer sur une seule œuvre d'un auteur, mais par exemple de partir d'une œuvre, d'ajouter une autre œuvre du même auteur puis une œuvre d'un auteur différent. Les critères à respecter se résument donc à ceci : varier les auteurs et choisir des thématiques différentes au sein d'un même genre.

La taille théorique du corpus est un autre critère à déterminer. Un corpus trop petit risque de conduire à une surreprésentation ou sous représentation de certains traits sémantiques. Ainsi, si l'étude effectuée est centrée sur un mot relativement rare et si la présence de ce mot est un critère de sélection des textes du corpus, le mot sur lequel l'étude est centrée sera surreprésenté. Par ailleurs, le corpus doit donner une image globale, donc sa taille doit être suffisamment importante. La référence suivante paraît pertinente : la méthode LSA s'appuie sur des corpus d'environ trois millions de mots. Signalons toutefois que notre étude ne requiert pas nécessairement une telle taille. En effet, les mots ne sont plus la référence mais les traits sémantiques. Or un mot correspond en moyenne à dix traits sémantiques si le TLFi est utilisé pour affecter les traits sémantiques (voir partie 4.3)). Nuançons toutefois : le recours au trait implique également une réduction du nombre de mots concernés. Certaines catégories de mots (déterminants, pronoms) sont en effet filtrées et n'apportent pas de trait sémantique. Les expériences menées ont reposé sur des corpus beaucoup plus petits pour des raisons techniques abordées au paragraphe 5.1.3.

Autre point fondamental : la structuration interne du corpus. Le décompte d'occurrences, qui initialise la représentation mathématique, repose sur le découpage du corpus. La structure interne justifiant le découpage ne doit donc pas être une maladresse mais une volonté sémantique de l'auteur. L'existence d'une structure simple, avec possibilité d'effectuer un seul découpage (en paragraphes en l'occurrence), a suffi à nos expériences. Néanmoins, la possibilité de structurer plus finement semble indispensable pour les évolutions ultérieures du modèle.

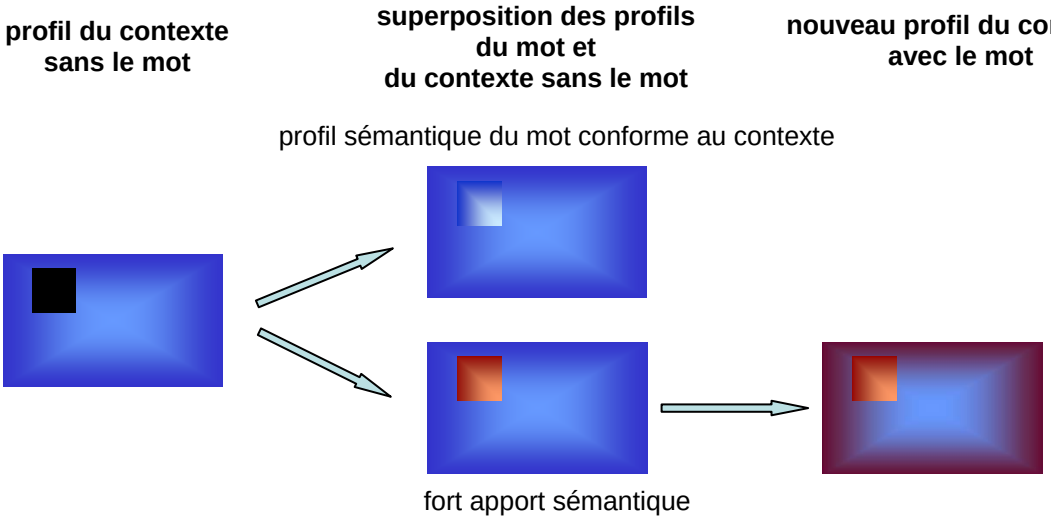
Pour conclure sur le corpus, rappelons qu'un corpus se constitue en fonction des objectifs fixés. Un choix d'homogénéité pour certains critères devrait s'accompagner d'une étude contrastive. Ainsi, les résultats d'une expérience sur un corpus de textes du XVIIe siècle devraient ensuite être confrontés à ceux d'un corpus du XXe siècle par exemple.

La sélection du mot demande elle aussi un examen attentif. Doit-il être polysémique ou monosémique[6] ? Quelle doit être la taille de son sémème ? Vaut-il mieux privilégier un mot rare ou un mot fréquent ? Voici quelques éléments de réponse. Avant tout, les questions précédentes sont liées : un mot dont l'usage est fréquent a généralement une plus grande richesse polysémique qu'un mot rare. Si un mot polysémique se voit affecter l'ensemble de ses traits sémantiques, toutes définitions confondues, la taille de son sémème est accrue. Deuxième élément de réponse : les critères de choix dépendent de l'expérience à mener. Si l'objectif est de faire émerger les traits sémantiques dominants, un mot monosémique convient. S'il est de structurer le sémème, la taille du sémème doit être suffisante pour qu'un réelle structuration apparaisse. Par exemple, le sémème de fragrance, constitué de deux sèmes, /odeur/ et /agréable/ ne sera pas approprié. S'il s'agit de désambiguïser[7], le choix d'un mot polysémique est indispensable. Par ailleurs, la sélection doit aussi prendre en compte les capacités humaines d'analyse : avoir une analyse humaine fine d'un sémème constitué de cent traits sémantiques sera beaucoup plus ardu que l'analyse d'un sémème de vingt traits, faute de vue d'ensemble.

[6] monsémique : qui n'a qu'un seul sens ; polysémique : qui peut prendre plusieurs sens
[7] désambiguïser : déterminer le sens approprié à un contexte donné d'un mot polysémique

Le choix du cotexte donne également matière à réflexion. Premier point à examiner : la taille du cotexte. L'unité de définition du cotexte et l'unité de découpage du corpus doivent-elles être identiques ? En particulier, supposons que le cotexte soit défini à partir d'une unité plus petite que l'unité de découpage ou à partir d'unités non multiples de celle-ci. Ce cas de figure peut de prime abord sembler plus problématique, voire aberrent. L'analogie suivante illustre le nœud du problème : alors que les observations qui ont permis de construire notre support ont été faites à l'œil nu, peut-on en tirer des conclusions au niveau microscopique ? Première remarque : les entités observées, à savoir les traits sémantiques, sont toujours les mêmes, elles ne sont donc pas transformées par le changement d'échelle. En revanche, la structure des entités observées change. Or les opérations effectuées sur le corpus ont précisément pour objectif de s'affranchir de la structure et de donner une représentation générale des relations entre traits sémantiques, à laquelle n'importe quelle structure pourra être appliquée. La sélection d'un cotexte dont l'unité de définition est quelconque se justifie donc. Signalons cependant que cette structure particulière doit se traduire, au niveau mathématique, par une fonction reflétant ses caractéristiques.

Par ailleurs, faut-il intégrer ou non le mot dans le cotexte ? Autrement dit, une fois sélectionné le cotexte centré le mot, faut-il en retirer le mot lui-même ? A nouveau, un tel procédé dépend des expériences à effectuer. Ainsi, la double approche, cotexte sans mot et cotexte avec mot, paraît appropriée pour mesurer d'une part la conformité d'un mot avec le cotexte dans lequel il apparaît, d'autre part ce qu'il apporte spécifiquement. L'illustration suivante peut s'avérer éclairante : le cotexte crée une toile sémantique, dans laquelle nous attendons qu'un terme avec un certain profil sémantique prenne place. Celui-ci peut avoir le profil attendu ou au contraire apporter quelque chose d'inattendu et, du même coup, rétroagir sur la toile du corpus pour la redessiner.



Une telle approche peut être, à terme, particulièrement intéressante pour l'étude de néologies (non développée dans le cadre de ce stage).

4.3) Pré-traitements

4.3.1) Découpage du corpus

Le découpage du corpus, indispensable pour constituer la matrice d'occurrences, pose des problèmes de deux ordres : l'échelle de découpage et la conservation de l'ordre des unités découpées.

4.3.1.1) Multiplicité des échelles sémantiques

Un corpus de textes est composé de nombreuses unités : sèmes, mots, mots composés (pomme de terre), syntagmes (regroupements de mots selon la structure grammaticale), collocations (poêle à frire, placard à balai), expressions figées ou semi figées (tuer le temps), phrases, paragraphes, chapitres, le textes… Ces unités constituent les candidats à un découpage, unités seules ou regroupées, comme par exemple les regroupements de mots, aussi appelés fenêtres de mots. Ces unités, qui entretiennent des relations d'inclusion illustrées ci-dessous, permettent de constituer un jeu d'échelle.

$$sème \subset mot \subset \left\{ \begin{array}{c} syntagme \subset phrase \subset paragraphe \\ fenêtre \quad de \quad mots \end{array} \right\} \subset texte \subset corpus$$

Les découpages possibles sont donc multiples, mais leur pertinence est variable en fonction de ce que l'on désire observer. Rappelons que notre étude est axée sur les occurrences et cooccurrences de traits sémantiques au sein de l'unité de découpage choisie. L'interaction entre un trait sémantique présent à la troisième page d'un roman et un trait à la cent cinquantième page a de fortes probabilités d'être moins significative qu'entre deux traits sémantiques distants d'au plus cent mots ; en revanche, qu'en est-il entre deux paragraphes ? Un trait sémantique présent en début d'un paragraphe est-il en plus étroite relation avec un trait en fin de ce même paragraphe ou un trait en fin du paragraphe précédent ? Une loi générale est difficile à dégager, d'autant plus que les études sur les sèmes sont presque inexistantes. En revanche, pour les mots, les avis des linguistes convergent sur certaines tendances, par exemple sur l'importance dominante des quatre termes placés en tête d'une définition. Or la place du mot correspond aussi à la place de son sémème, donc étendre aux sèmes les résultats obtenus sur les mots ne semble pas, a priori, complètement aberrent.

En premier lieu, les regroupements de paragraphes, chapitres, textes ou un corpus tout entier constituent des unités trop importantes. Les fenêtres de mots ont également une pertinence contestable. Elles soulèvent en particulier le problème de la fragmentation d'unités sémantiques : affectation de deux parties de paragraphe à des unités différentes, phrases coupées en plein milieu, etc. Néanmoins, elles pourraient avoir leur intérêt pour la définition d'une borne supérieure. Ainsi, couplées à un découpage en paragraphe, elles permettraient d'éviter des unités de découpage trop importantes. De même, elles constituent un recours en cas d'impossibilité technique d'accéder aux autres découpages (voir paragraphe sur les limites de Frantext). L'accès aux autres découpages n'est au demeurant pas insoluble, mais peut exiger le développement d'outils de découpage, parfois coûteux en temps. Les fenêtres de mots ou regroupements d'autres unités ont un autre avantage : elles permettent d'utiliser des fenêtres glissantes, c'est-à-dire un ensemble de fenêtres qui ne forment pas une partition mais se chevauchent et s'obtiennent par réitération d'une translation. Le syntagme est une unité sémantique pertinente mais trop petite parce qu'on désire observer un phénomène sémantique. Le choix de la phrase est lui aussi contesté pour les mêmes raisons. Les linguistes s'accordent en revanche sur la pertinence du paragraphe comme unité sémantique.

Le découpage idéal semble donc difficile à obtenir : plusieurs unités sémantiques font sens et aucun découpage unique ne peut prendre en compte cette arborescence de structures. Plutôt que de déterminer un découpage simple, ne serait-il pas judicieux de recourir à des découpages multiples ? Un découpage multiple permettrait d'intégrer les informations apportées par les syntagmes, les phrases et les paragraphes par exemple. Cependant, ce découpage multiple soulève de nouveaux problèmes : comment intégrer ces différents découpages ? Le traitement de chaque découpage doit-il être identique ? Différent ? Auquel donner la prépondérance ? Une solution consisterait à générer pour chaque découpage l'image du corpus par les mêmes transformations mathématiques, puis à faire la moyenne éventuellement pondérée des valeurs obtenues. Mais une nouvelle question se pose : est-il judicieux de synthétiser les informations apportées par chaque découpage en aval des transformations mathématiques ? Un travail en amont ne conviendrait-il pas mieux ? Ces questions sont importantes et n'ont pas été résolues. L'approche adoptée est la suivante : le corpus a été soumis à un découpage simple, procédé certes grossier mais cohérent, moins complexe et réalisable avec les outils existants. Ce modèle grossier des premiers pas appelle naturellement à être affiné à terme. Il devra s'appuyer sur les réflexions mentionnées ci-dessus.

4.3.1.2) Ordre : conservation ou non ?

Le découpage en paragraphes exclusivement, sans prendre en compte les unités supérieures et inférieures, et le traitement matriciel ramenant à un espace de traits sémantiques affranchis des découpages (matrice de cooccurrences, voir 4.4.1.1) font perdre toute notion d'ordre. A titre d'illustration, le corpus peut être vu comme un puzzle. Les morceaux de puzzle sont les unités de découpage (les paragraphes dans les études menées) et sont mélangés. Chacun de ces morceaux contient de nombreuses unités, les traits sémantiques, affectés d'un poids mais sans aucun agencement les uns par rapport aux autres. Les liens entre traits sémantiques ne dépendent donc que des cooccurrences au sein d'un même morceau de puzzle.

Le modèle choisi, qui a certes sa cohérence, peut toutefois paraître quelque peu brutal. En effet, pour des êtres humains, l'ordre semble intervenir en permanence et paraît fondamental. La lecture d'un texte est linéaire, la syntaxe repose elle aussi sur une relation d'ordre, des études linguistiques ont souligné que l'influence des éléments en tête de paragraphe (respectivement de phrase) est dominante au sein du paragraphe (resp. de la phrase) textes.

L'approche matricielle mise en œuvre au cours de ce stage, de même que la plateforme Sémy d'annotation en traits sémantiques, ne prennent pas en compte l'ordre. Nous avons opté pour un modèle plus simple mais robuste pour les premiers pas, conscients des imperfections de ce modèle. Ce modèle appelle à être affiné et complexifié par la suite. Les réflexions qui suivent s'attaquent au sujet, de manière plus ou moins approfondie selon les cas, et ouvrent de perspectives sur une intégration ultérieure de l'ordre dans le modèle.

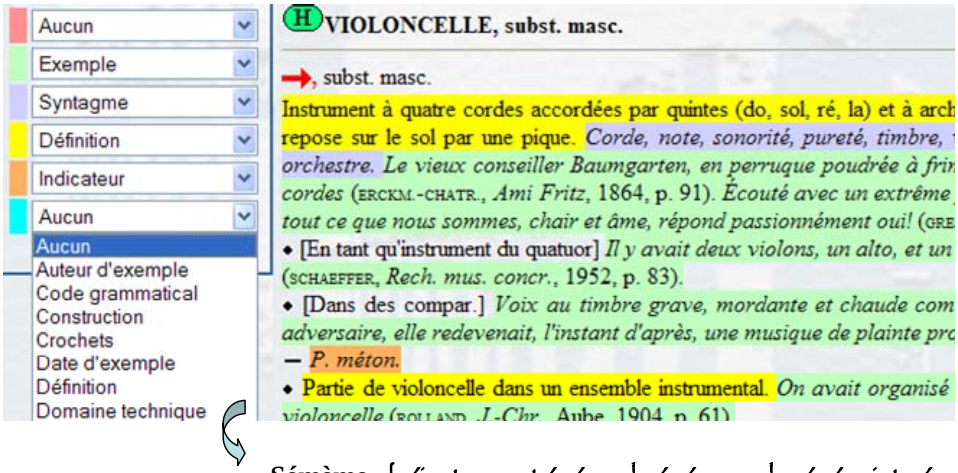
L'ordre peut être intégré à différents niveaux. Tout d'abord, pour un découpage simple, l'ordre pourrait être pris en compte lors des transformations matricielles. En particulier lors du passage de la matrice d'occurrences, où les paragraphes sont encore ordonnés (première colonne : premier paragraphe ; deuxième colonne : deuxième paragraphe,…), à la matrice de cooccurrences (lignes et colonnes correspondent à des traits sémantiques), la structure des paragraphes peut être en partie préservée par une transformation mathématique adéquate ; pour ce faire, les opérations sur les coefficients doivent s'appuyer sur les écarts d'indices de colonnes $\left|j_1 - j_2\right|$, qui est une distance linéaire. Une autre approche consisterait à faire des développements mathématiques parallèles, sous forme matricielle par exemple (constitution d'une autre matrice de cooccurrences) ne prenant en compte que les distances linéaires. Celles-ci dépendraient de l'unité de découpage choisie, à savoir le paragraphe. Par exemple, les traits sémantiques distants de n paragraphe(s) auraient un coefficient de cooccurrence de f(n) dans la matrice de cooccurrences. La distance linéaire pourrait également être plus complexe et prendre en compte les multiples unités sémantiques : deux traits distants de m paragraphes, n phrases et p mots auraient une distance d fonction de m, n et p. Une des difficultés majeures réside dans le choix d'une fonction adaptée, susceptible de refléter de manière juste les phénomènes linguistiques. Une autre approche est celle mentionnée au paragraphe précédent : elle reposerait sur des transformations menées en parallèle sur les différents découpages et une synthèse des objets mathématiques finals. Dernière approche envisagée : travailler sur des voisinages glissants. Le découpage ne se ferait plus en unités disjointes mais en unités qui se recouvreraient. Ainsi, dans une optique de paragraphes, le premier élément correspondrait aux paragraphes 1, 2 et 3, les second aux paragraphes 2, 3 et 4, … Toutefois, cette solution pose un problème non négligeable : celui des effets de bord. Enfin, une dernière remarque s'impose sur l'ordre linéaire : celui-ci est certes valable à échelle suffisamment grande mais peut poser problème à échelle trop fine (inférieure à la phrase), où la syntaxe influe fortement. Faudrait-il alors substituer à l'ordre linéaire un ordre logique ? Et, si cela est théoriquement possible, les outils capables d'établir cet ordre existent-ils et sont-ils adaptables au modèle ?

4.3.2) Affectation des traits sémantiques

4.3.2.1) Source des traits sémantiques

L'affectation des traits sémantiques repose sur la plateforme d'annotation Sémy. Celle-ci a pour base le TLFi et s'appuie sur les regroupements morphologiques de [Ramdani, 2007].

A chaque unité textuelle est associé un ensemble de sèmes. Ceux-ci sont obtenus par le TLFi qui contient le sémème de chaque mot retenu dans l'unité textuelle. Les traits sémantiques sont puisés dans les entrées associées à chaque mot. Seule une partie des entrées a été retenue : les définitions. Les autres rubriques (exemples, leurs auteurs, dates et titres, syntagmes, domaines techniques, synonymes,…) ont été éliminées.



Sémème : {/instrument/,/corde/,/accorder/,/quinte/, ...}

Remarquons néanmoins que les domaines auraient eu leur pertinence. Cette extension devrait être intégrée à terme par le concepteur de Sémy, la plateforme d'annotation en traits sémantiques qui permet la sélection des traits appropriés (catégorie grammaticale, rubrique du TLFi, etc). La sélection d'une seule rubrique a un avantage majeur : elle permet de limiter le nombre de traits sémantiques affectés à un mot, donc réduit la combinatoire lors de calculs ultérieurs ainsi que le 'bruit', c'est-à-dire la proportion de traits sémantiques non pertinents affectés à un mot.

Deux autres extensions méritent d'être prises en considération : l'adjonction du mot lui-même dans son sémème et l'enrichissement par sémème inverse. Le premier point soulève la question de la réflexivité : un mot fait-il partie de son propre sémème ? Ne fait-il partie de son propre sémème que s'il est repris dans sa définition ? L'exemple suivant met en lumière un argument en faveur de la réflexivité : si le mot **sable** a le trait sémantique /jaune/, et le mot **jaune** n'a pas /jaune/ comme trait sémantique, l'unité de sens /jaune/ sera plus présente dans « Il marche dans le sable » que « Tout son intérieur était jaune : chaises jaunes, canapé jaune, mur jaune ». Dans les expériences menées, la réflexivité n'a pas été systématisée pour des raisons techniques. Si le modèle est développé, cette faille devra être comblée. La seconde extension est celle du sémème inverse. Le sémème inverse est un sémème obtenu par la démarche inverse de celle effectuée : les traits sémantiques du sémème étaient les termes de la définition du mot de référence ; les traits sémantiques du sémème inverse sont les termes dont le mot fait partie de la définition. Prenons un exemple : le mot **vibrato**. La définition du TLFi est la suivante : *« technique d'interprétation destinée à rendre un son plus expressif en faisant varier légèrement et très rapidement sa hauteur »*. Le sémème correspondant est : {/technique/, /interprétation/, /son/, /plus/, /expressif/, /faire/, /varier/, /légèrement/, /très/, /rapidement/, /hauteur/}. Par ailleurs, **vibrato** apparaît dans les définitions de **sifflet, tremblant, vibrant, vibrer** et **voix**. Le sémème inverse est donc : {/trembler/, /vibrer/, /sifflet/, /voix/}. Le sémème total, constitué du sémème 'direct' et du sémème inverse, est enrichi et paraît plus pertinent. Toutefois, le sémème pose également un certaines nombre de problèmes : pour les mots polysémiques, le bruit sera amplifié. Le mot **palais** se verra enrichir aussi bien de traits sémantiques comme /élyséen/, /alcazar/ ou /évêché/ que d'/arrière-gorge/, /amygdale/ ou /voyelle/. On imagine l'effet dans un corpus de contes parlant de princes et de leur palais : les traits indésirables seront multipliés. Au problème de la pertinence s'ajoute celui de la quantité. Par exemple, le sémème

de rose (nom et adjectif réunis) comporte un peu plus de cent sèmes. Or **rose** apparaît dans 160 définitions. Le sémème total serait donc plus que doublé. Ajoutons un troisième bémol pour les mots polysémiques : le TLFi est structuré en définitions et la sélection de la définition appropriée est techniquement possible. En revanche, lorsqu'un mot apparaît dans une définition, aucune indication n'est enregistrée sur le sens ad hoc. Autrement dit, traiter la polysémie indésirable semble possible avec le sémème direct mais d'un autre niveau de complexité dans le cas du sémème inverse. En somme, le sémème inverse peut s'avérer précieux mais doit être manipulé avec précaution. Une piste à explorer consisterait à définir le type de mots à enrichir par sémème inverse. Les candidats pourraient être des mots monosémiques, d'une catégorie grammaticale déterminée (nom par exemple) et dont le sémème direct n'excèderait pas une certaine taille.

4.3.2.2) Filtrage et regroupement des sèmes

Certains traits sémantiques sont filtrés. Les éléments conservés sont les traits sémantiquement pleins, d'où la sélection des catégories grammaticales suivantes : noms, verbes, adjectifs et adverbes. Les autres catégories (pronoms, déterminants, …) sont éliminées. De même, les éléments métalinguistiques hérités des définitions et convertis en sèmes (par ex., « qui n'a pas » devient un sème /absence de/, « qui est en rapport avec » ou « qui est caractérisé par » devient /caractéristique de/, …) sont retirés des listes.

La sélection effectuée dans les expériences réalisées a toutefois mis en lumière des faiblesses de cette sélection. Ainsi, la pertinence des adverbes est contestable. Par exemple, le sémème du mot **pollen** est constitué de dix-huit traits sémantiques. Parmi eux, deux adverbes : /très/ et /généralement/. Le manque de pertinence de ces adverbes est manifeste : le pollen n'évoque ni la généralité, ni l'intensité. Cependant, le retrait complet des adverbes est-il une solution adéquate ? Reprenons l'exemple de **vibrato** : supprimer les sèmes /rapidement/ et /légèrement/, sémantiquement riches et pertinent, priverait le sémème d'éléments fondamentaux. Le traitement des adverbes nécessiterait une solution panachée : la suppression d'une liste d'entre eux soit par étude statistique sur les définitions du TLFi (élimination des adverbes dont le taux de présence dépasse un certain seuil), soit par étude linguistique du même ordre que celle sur les métasèmes. D'autres points méritent considération : que faire des verbes tels « être », « avoir », « faire », « tenir » ? Certes, leur présence parasite souvent le sémème, mais leur apport est indispensable dans certains cas, comme dans les sémèmes de **détenir** (/avoir/), **posséder** (/avoir/) ou **serrer** (/tenir/).

Après la sélection s'impose un regroupement de traits sémantiques. Celui-ci s'opère par lemmatisation. La légitimité de la lemmatisation semble a priori évidente : l'essentiel du contenu sémantique d'un verbe semble identique quels que soient la personne, le nombre, le mode et le temps auquel il est employé ; les apports sémantiques de la conjugaison paraissent extérieurs au verbe, d'une source dissociée, indépendante du cœur sémantique du verbe. Soulignons toutefois que, dans des analyses linguistiques fines, ce caractère d'évidence est mis à mal. Brève illustration, reprise de [Bourion, 2001] : dans les romans du XIXe siècle, **pieds** et **pied** n'avaient pas la même coloration sémantique : les traits sémantiques /fin/, /petit/ et /gracieux/ étaient beaucoup plus rattachés au pluriel **pieds** qu'au singulier. Puis s'ajoute à la lemmatisation un regroupement en familles. Celles-ci sont les familles morphologiques décrites au paragraphe 3.3.3. A nouveau, ces regroupements entraînent une moindre finesse de sens. Examinons le cas particulier du sème /nature/ : celui-ci est regroupé avec /naturaliser/, /naturiste/, /naturel/ ou encore /dénaturé/. Les sèmes évoqués sont loin d'être synonymes. Pourtant, ils ont effectivement un point commun, sémantiquement parlant. Ces regroupements présentent un autre avantage : ils évitent une distribution de sens trop dispersée. Sur le plan mathématique, ils amoindrissent les problèmes liés aux matrices creuses.

En somme, le bien-fondé de la lemmatisation et des regroupements repose principalement sur l'argument suivant : la démarche adoptée est une démarche de modélisation. Celle-ci est une simplification du réel, elle repose sur des caractères généraux et réguliers. Un modèle capable de

rendre compte de toutes les subtilités de la langue ne serait plus un modèle... ou serait le modèle parfait, assassin de la liberté humaine à travers le langage.

4.3.3) Pondération des traits sémantiques

Avant toute opération mathématique se posent deux questions fondamentales : faut-il affecter une pondération aux traits sémantiques d'une définition avant tout traitement ? Le cas échéant, sur quels critères doit reposer cette pondération ?

La première question est en réalité un faux choix : ne pas pondérer entraîne implicitement la sélection d'une pondération par défaut. Celle-ci est une pondération présence/absence : le poids 1 est affecté au trait présent dans la définition du mot, quel que soit son nombre d'occurrences dans la définition, 0 au trait absent.

Les pondérations possibles recouvrent des aspects multiples. Première interrogation : quelle poids donner aux sèmes apparaissant plusieurs fois dans une définition ? Les apparitions doivent-elles être comptées de façon simple ou multiple ? Argument pour un décompte multiple : un terme qui apparaît plusieurs fois dans une définition peut être considéré comme plus significatif que d'autres. Prenons l'exemple du mot thé. Les traits sémantiques apparaissant plusieurs fois dans la définition sont : /feuille/ (6 occurrences), /boisson - boire/ (6), /thé/ (4), /servir/ (3), /préparer - préparation/ (3), /bourgeon/ (2), /plante/ (2), /infusion/ (2), /autre/ (2), /mode/ (2). Ces traits sémantiques semblent, pour la plupart, particulièrement caractéristiques du thé et une pondération proportionnelle au nombre d'occurrences ne semblerait pas aberrante. Cependant, les traits /autre/ et /mode/ mettent en lumière un autre problème : les traits sémantiques trop fréquents ou mots 'passe-partout' risquent de voir leur pondération démultipliée, alors qu'on souhaiterait au contraire les éliminer. Pour trancher la question, une étude rigoureuse, à la fois statistique et linguistique, des définitions du TLF s'imposerait et la pondération devrait probablement intégrer la fréquence des termes dans la langue française. Je n'ai pas axé mes efforts sur une telle étude, purement dictionnairique, puisque mon sujet était centré sur le contexte. Mon choix a été de garder la pondération 'présence – absence', indépendamment de l'ordre de multiplicité.

Deuxième question à aborder : celle de la normalisation. Considérons deux mots. Le sémème du premier comporte dix sèmes, celui du second quatre-vingt. Faut-il alors accorder un poids de 1/10 aux sèmes du premier mot et de 1/80 à ceux du second, ou le même poids à tous les sèmes, sachant que le poids du mot sera de 10 contre 80, du moins si le poids du mot est égal à la somme des poids de ses sèmes ? Grossièrement, il s'agit de choisir entre une égalité entre mots et une égalité entre sèmes. Les deux approches se défendent. Toutefois, mon approche infra-lexicale, détachée du niveau des mots, m'incite à privilégier une égalité entre sèmes, donc un retour à la pondération basique.

Une troisième pondération pourrait dépendre de la rubrique dans laquelle un trait sémantique apparaît. Rappelons que les traits sélectionnés appartiennent tous à la définition. Cependant, il serait envisageable d'enrichir le sémème avec les sèmes présents dans les exemples ou encore les synonymes. Selon la rubrique (définition, exemple,…) le sème pourrait se voir affecter un certain poids (1 pour la définition, ½ pour les autres rubriques par exemple). Mais au problème de la pondération adéquate s'ajoute un problème déjà mentionné en 4.3.2.1 : adjoindre de nouvelles rubriques entraîne une multiplication des sèmes du sémème.

Un autre mode de pondération à creuser s'appuierait sur la position dans la définition. Par position, on peut entendre position linéaire (nième mot de la définition, ou encore mot présent dans la $k^{ième}$ fraction) ou position syntaxique. Une telle pondération peut paraît prometteuse. Cependant, elle soulève de nouveaux obstacles : comment pondérer ? Par décroissance linéaire, logarithmique, exponentielle ? Une fois l'ordre syntaxique déterminé, quelles règles appliquer ? Le choix d'un modèle approprié mériterait une réflexion soignée, que certaines études linguistiques permettent déjà d'alimenter.

4.4) Traitements mathématiques

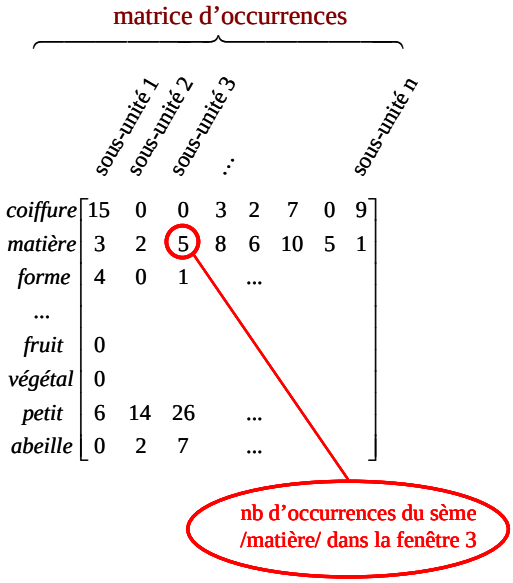
4.4.1) Matrice du corpus : du nombre d'occurrences à la significativité des cooccurrences

4.4.1.1) Point de départ : décompte des occurrences

Les briques de base qui permettent de faire le passage des traits sémantiques, données qualitatives, à une représentation quantitative sont les suivantes :

- la présence / absence (indicateur 1 / 0)
- le nombre d'occurrences, c'est-à-dire le nombre d'apparitions d'un trait sémantique ; il peut être absolu (sur l'ensemble du corpus) ou relatif (par unité de découpage)
- le cardinal du corpus (en traits sémantiques)
- le cardinal de chaque sous-unité du corpus
- le nombre de sous-unités

Ces éléments permettent de générer une matrice d'occurrences. Cette matrice a pour lignes les traits sémantiques du corpus, pour colonnes les sous-unités du corpus et pour entrées le nombre d'occurrences d'un trait sémantique dans une sous-unité du corpus.



L'indicateur choisi, à savoir le nombre d'occurrences, pourrait être affecté d'une pondération comme indiqué en partie 4.3, mais le choix effectué en accord avec mon équipe de travail a été de démarrer la réflexion et les expérimentations avec un corpus le plus simple et le plus robuste possible, à affiner et complexifier dans des développements ultérieurs.

4.4.1.2) Transformations matricielles

L'étape suivante consiste à transformer la matrice d'occurrences du corpus par diverses opérations. Ces opérations doivent permettre d'observer les résultats de phénomènes linguistiques de différents types, décrits ci-dessous. Elles ont deux vocations : générer des coefficients qui, à l'issue des opérations, reflètent la significativité des traits sémantiques ; se ramener à un espace de cooccurrences et non plus d'occurrences, dans lequel on considère les traits sémantiques non plus relativement à un texte découpé mais relativement à eux-mêmes. Ce deuxième objectif est étroitement corrélé avec la notion de forme sémantique : les traits sémantiques sont considérés à travers leurs relations et leurs regroupements, et non plus de manière indépendante. Cette nouvelle structure, bâtie sur les cooccurrences, est favorable aux jeux d'échelle et favorise les expériences sur des cotextes variés.

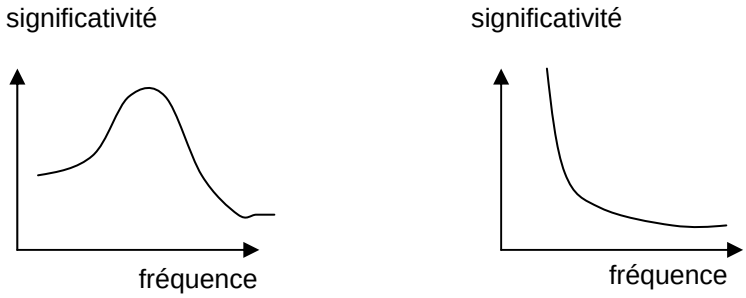
4.4.1.2.1) Fréquence et significativité : dans le sillage de Zipf

Le premier type de transformations effectuées tente de refléter le principe suivant, exprimé dans un premier temps au niveau lexical, puis à transposer au niveau infra-lexical : les mots très fréquents sont peu significatifs ; les mots rares ont une forte significativité dans le cotexte proche mais sont inhibés par l'effet de masse à plus grande échelle. Comme exemple, citons *être*, 3^{e} mot le plus fréquent de la langue française d'après le [Dictionnaire des fréquences, 1971], et *pouvoir*, au 42^{e} rang : l'apport sémantique de tels termes sera, en règle général, très faible.

Soulignons que la loi de Zipf ne prend pas en compte tous les phénomènes, en particulier pour des termes rares : la rareté ou l'absence d'un terme peut être d'une extrême significativité. D'après l'exemple de [Valette, 2006c], dans le corpus des 400 conférences du linguiste Gustave Guillaume, le mot « mécanisable » n'apparaît qu'une fois, alors que les dérivés de « mécanique » sont extrêmement fréquents. Or « mécanisable » est un néologisme inventé dans le contexte cybernétique. Son absence est en fait un moyen pour Gustave Guillaume de cacher ses sources qu'il ne cite d'ailleurs jamais. La représentation simplificatrice évoquée ci-dessus a cependant été jugée acceptable, au moins temporairement, au sein de l'équipe à composantes transverses dans laquelle je travaille.

Puis l'hypothèse suivante a été faite : ce phénomène, observable au niveau lexical, se reproduit au niveau infra-lexical, pour les traits sémantiques.

L'allure d'une courbe représentative de la significativité en fonction de la fréquence d'un trait sémantique pourrait donc être une des deux suivantes :



Les lois auxquelles ce comportement général s'applique sont la loi de Zipf, la méthode tf-idf ou encore l'entropie, mentionnées au 3.2.1, lois d'ailleurs en relation les unes avec les autres. Dans les expériences menées, la méthode tf-idf a été appliquée.

Quelques questions, non résolues, méritent d'être mentionnées : cette loi est-elle valable à toutes les échelles ? Pour un découpage en paragraphes, doit-elle être appliquée de manière inter- ou intra-paragraphes ? Si elle n'est plus valable à partir d'un certain seuil, quel est ce seuil ? Comment le déterminer ?

4.4.1.2.2) Repérage de la surreprésentation et sous-représentation

Autre phénomène en jeu : le taux de présence 'anormal', au sens statistique, d'un trait sémantique. Plus précisément, le nombre de cooccurrences observé entre deux traits sémantiques (resp. le nombre d'occurrences d'un trait dans un paragraphe donné) peut être considéré comme fruit du hasard ou significatif, connaissant le nombre de cooccurrences total de chacun des traits (resp. le nombre d'occurrences total du trait et le nombre total de traits dans le paragraphe considéré).

De manière plus formalisée et dans une optique statistique, nous pourrions reformuler cette assertion ainsi : considérons deux traits sémantiques. Sous l'hypothèse H0 d'indépendance de 2 traits (resp. du trait et du paragraphe), la probabilité que le nombre de cooccurrences (resp. occurrences) soit dans l'intervalle [a, b] est de n%. Si le nombre de cooccurrences (resp. occurrences) est hors de cet intervalle, la cooccurrence est considérée comme significative, les traits sont donc corrélés.

Le repérage de cooccurrences significatives s'appuie donc sur des méthodes statistiques dont l'objectif est de quantifier la significativité. Les méthodes repérées correspondant à cette démarche sont celle employée par [Victorri, 2005] qui s'inspire de la méthode du χ^2 et celle employée par Mauceri reposant sur le test de Fisher, avec la formule de filtrage des cooccurrences significatives. La méthode mise en œuvre dans les expériences est semblable à celle employée par Victorri, bien que celle de Mauceri mérite également d'être implémentée. Faute du temps nécessaire, seule celle de Victorri a été appliquée.

4.4.1.2.3) Psycho-linguistique et gestion de la multiplicité de sens

La troisième ligne de force des applications mathématiques choisies repose sur des critères psycho-linguistiques. Sommairement, le principe est le suivant : l'esprit humain ne peut gérer une trop grande

multiplicité sémantique, il se ramène donc à de grandes lignes, qu'on peut concevoir comme des directions principales sur lesquelles il projette le reste. Cette formulation est certes schématique et ne prétend donner qu'une approche grossière du problème

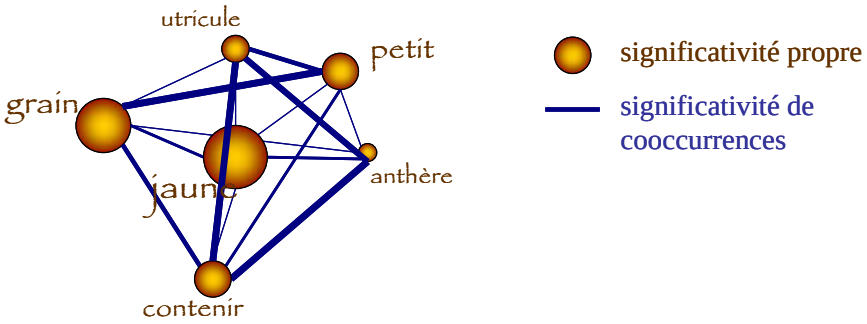
Une approche plus mathématique consiste à se représenter les traits sémantiques comme un espace multidimensionnel. Une projection est alors effectuée selon certains axes, en l'occurrence les axes principaux, à partir de l'ensemble des données disponibles.

Les travaux mathématiques dans la veine de cette conception sont l'analyse par composantes principales et l'analyse sémantique latente (LSA) qui découle de l'ACP. La transformation mathématique appliquée dans les expériences a tenté de reproduire les opérations mathématiques pratiquées pour LSA.

4.4.1.2.4) Des occurrences aux cooccurrences

Une autre étape dans les transformations mathématiques consiste à passer d'une matrice d'occurrences, qui reflète la présence ou le poids des traits sémantiques en fonction de leur sous-unité de corpus d'apparition, à une matrice de cooccurrences, où les poids ou présence des traits sémantiques s'expriment en fonction des autres traits sémantiques.

Cette transformation permet de s'affranchir du découpage et d'appliquer d'autres découpages. Elle est indispensable, ou du moins d'un grand secours pour passer d'observations globales à des observations locales. Signalons cependant que les informations apportées par le découpage du corpus ne sont pas complètement perdues : elles seront en partie contenues dans les coefficients de la nouvelle matrice, bien qu'il y ait tout de même perte d'information. La génération d'une matrice de cooccurrences répond également à une autre logique : la comparaison de traits sémantiques. Par cette transformation, les traits sémantiques sont ramenés les uns aux autres. Ajoutons un autre élément : un trait sémantique peut être vu comme un élément affecté d'un poids propre, lié à sa seule présence, et d'un poids de cooccurrence, lié à la présence des traits sémantiques avec lesquels il apparaît. Voici une représentation imagée du phénomène en jeu : les traits sémantiques peuvent être considérés comme des atomes de masse ou dimension variables (leur significativité propre) et sont reliés deux à deux à tous les autres traits sémantiques par des liaisons de force variable (la significativité de cooccurrence).



Différentes transformations peuvent être appliquées pour obtenir la matrice de cooccurrences. Pour appréhender les opérations effectuées, représentons-nous la matrice d'occurrences comme un ensemble de vecteurs (les lignes) dans l'espace à p dimensions des sous-unités du corpus (paragraphes en l'occurrence), p étant le nombre de sous-unités. Les transformations possibles sont :

- la multiplication de la matrice d'occurrences par sa transposée : le coefficient (i,j) de la nouvelle matrice correspond au produit scalaire entre le vecteur ligne (trait sémantique) i et le vecteur ligne (trait sémantique) j. Intuitivement, cela signifie que le poids de cooccurrences de deux trait croît avec la significativité de chaque trait et avec la similarité de distribution (diminution de l'angle entre les deux vecteurs).

- par calcul du cosinus entre deux lignes (coefficient en position (i,j) : cosinus entre les vecteurs des lignes i et j de la matrice d'occurrences) : dans ce cas, la norme des vecteurs (significativité totale) ne joue plus, seul compte la similarité de distribution des sèmes, autrement dit l'angle entre vecteurs. L'intérêt du cosinus est que ses valeurs sont comprise entre -1 et 1, ce qui facilite l'interprétation, plus intuitive.

- par tout autre mesure de distance entre deux vecteurs, effectuée sur tous les couples (i,j) de vecteurs-lignes de la matrice d'occurrences.

Se ramener d'une distance à un poids, coefficient d'affinité ou encore de significativité, est aisé : prendre l'inverse de la distance (en traitant le cas du zéro) ou, dans le cas d'une fonction bornée en valeur absolue par M, retrancher à M la distance sont des solutions relativement triviales.

Le produit de la matrice par sa transposée et le cosinus ont été appliqués dans les expériences. L'exploitation d'autres mesures reste une voie à explorer.

4.4.1.2.5) Ordre d'application des transformations

Les différents modèles présentés ont tous leur cohérence et peuvent même paraître complémentaires. Cependant, si l'on choisit d'appliquer plusieurs transformations, dans quel ordre appliquer celles-ci ? En effet, ces transformations ne sont pas commutatives et imposent de se représenter la manière dont les traits sémantiques sont affectés par les transformations.

L'exercice d'aller-retour entre plan mathématique et plan linguistique a constitué un obstacle réel, attaqué mais non surmonté. Plutôt que de chercher à obtenir une vision théorique limpide avant d'effectuer les expériences, j'ai choisi de m'appuyer sur les expériences pour qu'elles confortent ou guident les intuitions théoriques. Le choix de la bonne transformation s'est appuyé sur différents essais où seule la succession de transformations variait, puis sur une analyse comparative des différents résultats obtenus.

Les cas de figure possibles sont l'application successive des transformations mathématiques (composition de fonctions) ou application séparée des transformations (application de fonctions différentes à la matrice de départ) puis synthèse des différents résultats obtenus (fonction de plusieurs matrices). Dans les expériences menées, la composition s'est bornée à deux, au maximum trois fonctions, la dernière étant, sauf dans un cas le passage de la matrice d'occurrences ou de cooccurrences (produit de la matrice par sa transposée ou cosinus). Les combinaisons de transformations appliquées sont détaillées en partie expérience (voir 5.2.2).

4.4.1.2.6) Interprétation du produit final

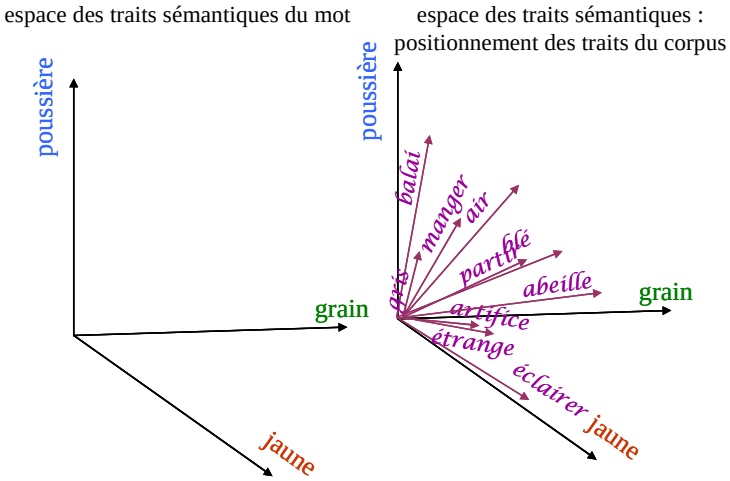
La matrice finale est une matrice symétrique, dont les lignes et les colonnes correspondent aux traits sémantiques du corpus. Les coefficients de la matrice sont les poids de significativité affectés à une paire de traits sémantiques. L'ensemble de la matrice est une représentation mathématique du corpus, une vision globale (c'est-à-dire à l'échelle du corpus) des unités sémantiques et de leurs interactions.

4.4.2) Du global au local : représentation du mot et de son cotexte

Une fois la représentation globale obtenue, nous souhaitons nous ramener à une étude locale : celle d'un mot dans un cotexte proche.

4.4.2.1) Le mot

La représentation du mot s'effectue à travers son sémème. Ce sémème peut être considéré comme un espace sémantique à n dimensions, où n est le nombre de sèmes composant le sémème. Par exemple, le mot *pollen*, composé de dix-huit traits sémantiques, peut se voir comme un espace à dix-huit dimensions : la dimension 'jaune', la dimension 'grain', la dimension 'poussière', …



Les dimensions sont équivalentes, autrement dit l'espace dans lequel on se place est isotrope dans notre cas. En effet, seule compte la présence / absence dans le sémème. Un autre choix aurait pu être fait, par exemple d'accorder la préséance aux sèmes présents de manière multiple dans la définition, ou à ceux présents en tête de définition. Un tel choix se serait traduit par une anisotropie de l'espace.

L'espace du mot correspond à un sous-espace de la matrice représentative du corpus : il s'agit de la projection orthogonale de la matrice du corpus sur l'espace des traits sémantiques du mot. Cette projection correspond à une sélection de colonnes (les lignes auraient aussi convenu, la matrice étant symétrique). Rappelons que les colonnes de la matrice correspondent chacune à un sème. Les colonnes sélectionnées sont celles qui correspondent aux sèmes présents dans le sémème du mot de référence.

représentation matricielle de
la sélection des traits du mot

	balai	*air*	*poussière*	…	*étrange*	*jaune*	*blé*	*grain*
balai	35	3	0	…	2	7	0	9
air	3	28	5		6	10	5	1
poussière	0	5	42		…		26	7
…	…				…			
balai	2	6						
jaune	7	10						
blé	0	5	26		…			
grain	9	1	7		…			

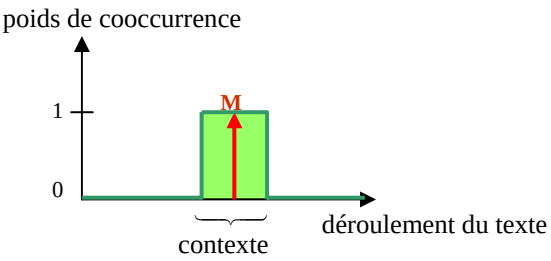
4.4.2.2) Le cotexte

Lors de l'étape précédente, l'ensemble des lignes de la matrice corpus-mot correspondait à un ensemble de vecteurs dans l'espace sémantique du mot.

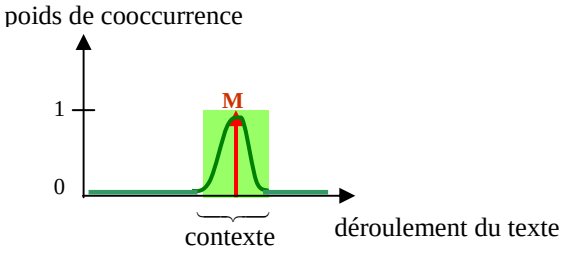
Pour se ramener au cotexte, le choix effectué est de ne garder que les vecteurs correspondant aux traits sémantiques présents dans le cotexte. Au niveau matriciel, cette opération correspond à la sélection d'une sous-matrice de la matrice corpus-mot : la matrice cotexte-mot. Ses colonnes sont les mêmes que celles de la précédente et ses lignes sont celles correspondant aux sèmes du cotexte. Pour compléter cette approche, reprenons la représentation atomique, avec des atomes-sèmes et des liens entre toutes les paires d'atomes. La constitution de la matrice corpus-mot correspond à une sélection de liens : seuls sont gardés les liens entre les sèmes du mot et les autres sèmes du corpus. La sélection du cotexte revient à supprimer les liens avec tous les atomes-sèmes différents des sèmes du cotexte.

Revenons enfin sur un problème mentionné en 4.2) au sujet de l'unité de définition du cotexte. Nous soulignions alors la possibilité de prendre un cotexte de taille quelconque, pourvu que cette sélection du cotexte se traduise au niveau mathématique par une fonction reflétant la structure du cotexte. Dans notre cas, la fonction appliquée est une simple fonction créneau. Elle pourrait prendre différentes formes, en particulier refléter le centrage sur le mot de référence.

Fonction créneau



Autre fonction reflétant le centrage sur le mot



M : mot de référence

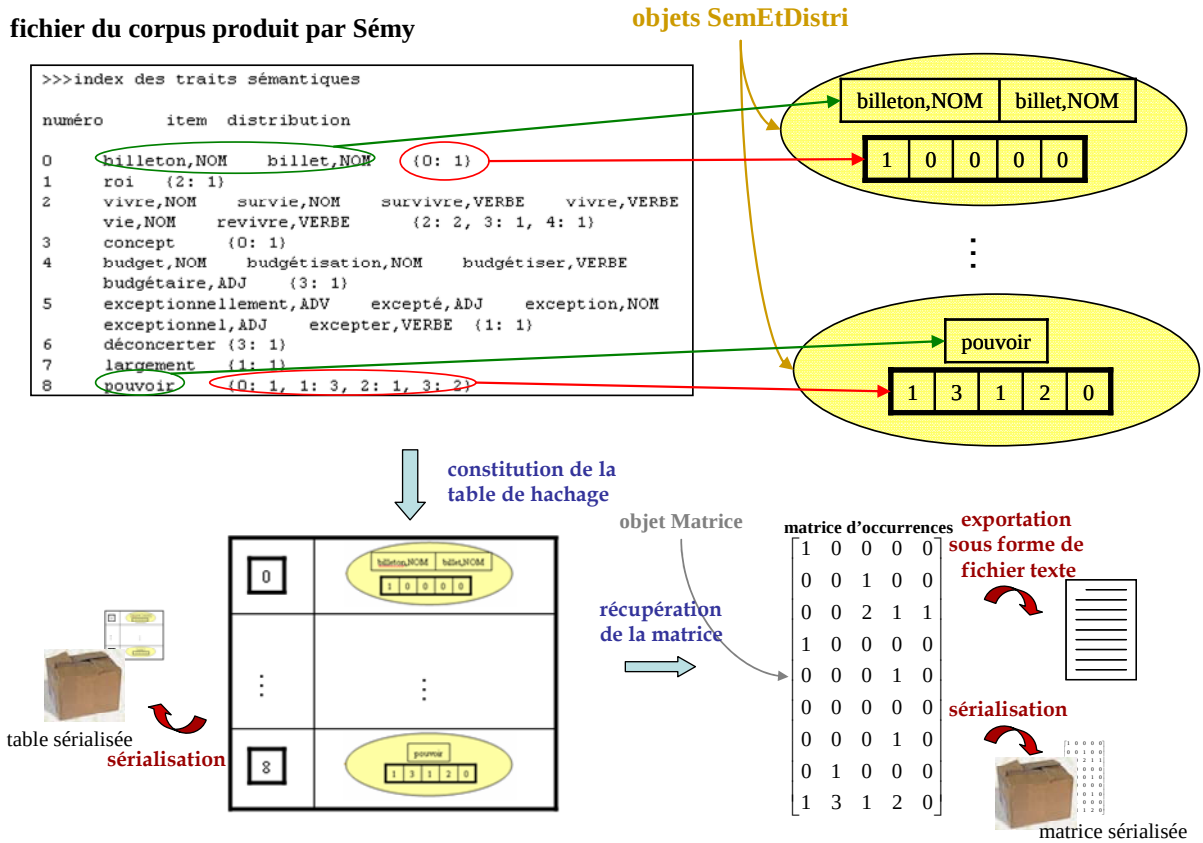
V) Expérimentations

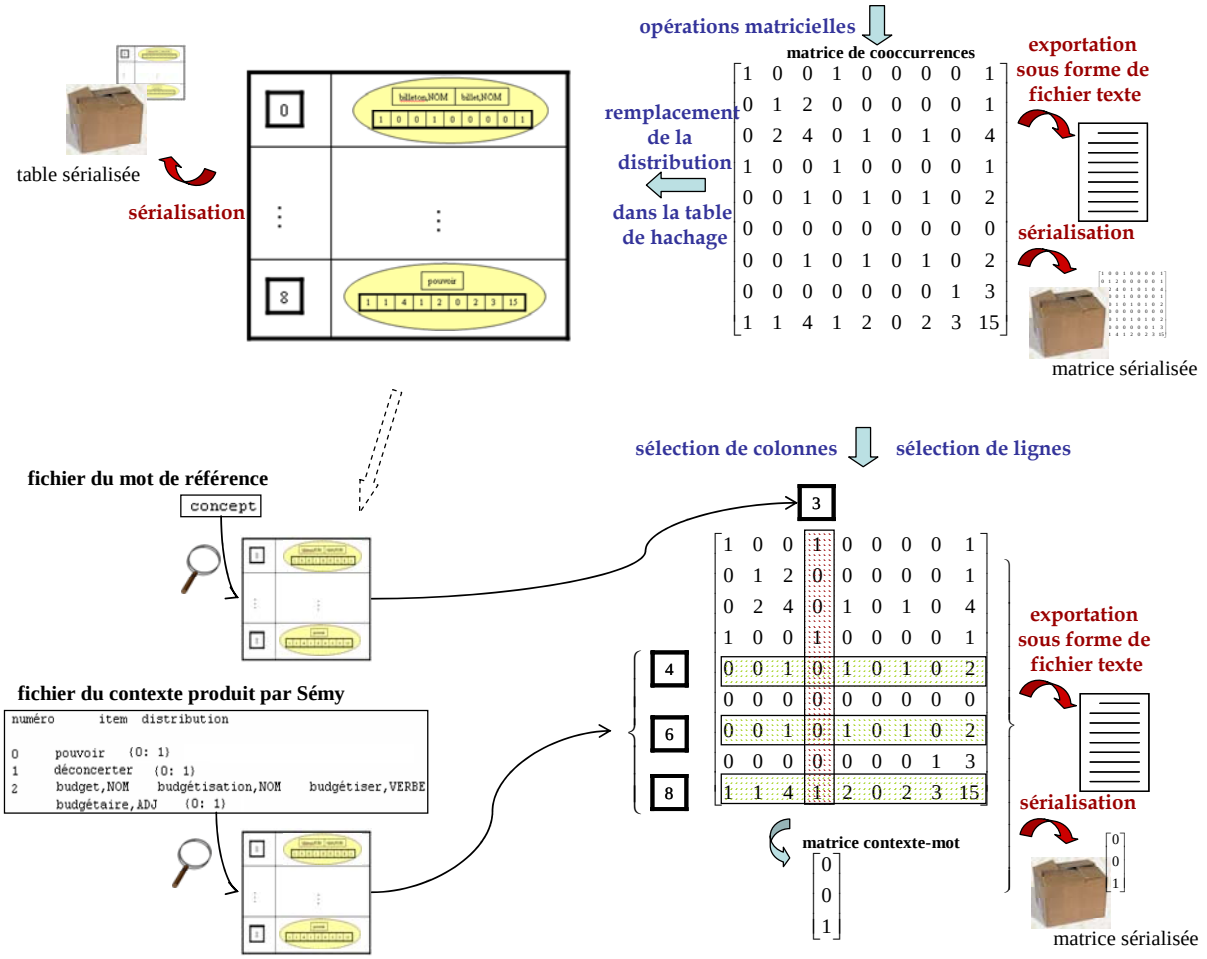
A la réflexion théorique sur les modèles adéquats a succédé une phase pratique, avec mise en place d'un outil expérimental (programme en Java). Cette phase pratique a consisté en la détermination et la réalisation d'une série de tests en fonction des objectifs fixés et des contraintes techniques, puis analyse des résultats obtenus.

5.1) Automatisation des transformations : programmation en Java

5.1.2) Architecture

Le programme informatique prend en entrée trois fichiers générés par le programme Sémy : un fichier correspondant aux traits sémantiques du corpus découpé en sous-unités et leur distribution ; un fichier correspondant aux traits sémantiques du cotexte ; un fichier correspondant aux traits sémantiques du mot. Il génère une matrice d'occurrences du corpus dont les lignes sont associées à des traits sémantiques et les colonnes aux unités de découpage, il effectue différentes transformations mathématiques (voir partie 4.4) sur cette matrice, en extrait l'image du mot puis du cotexte et retourne de fichiers contenant les résultats obtenus ainsi que des fichiers contenant toute l'information sur les objets créés. Le schéma suivant illustre le fonctionnement de ce programme et sera suivi d'explications plus détaillées :





Pour chaque ligne du fichier des traits sémantiques du corpus, le programme crée, à partir d'une classe spécifique (SemEtDistri), un objet qui contient deux informations : la famille de traits sémantiques, stockée dans une liste de type ArrayList, et la distribution, stockée dans un tableau à p colonnes, où p est le nombre de sous-unités du corpus.
Il génère ensuite une table de hachage, c'est-à-dire une structure qui relie deux types d'objets : les valeurs et les clés par lesquelles on accède aux valeurs. Dans notre cas, les clés sont les indices des traits sémantiques et correspondront par la suite aux indices de ligne des matrices ; les valeurs sont les objets SemEtDistri.
A la table de hachage est ensuite associé un objet Matrice. Le vecteur de la $(i+1)^{ème}$ ligne correspond à la distribution (tableau comportant le nombre d'occurrences) de la famille de sèmes d'indice i.
Différentes méthodes sont ensuite appliquées aux matrices, chacune correspondant à une transformation mathématique (tf-idf, LSA, transformation inspirée du χ^2 semblable à celle de Victorri, produit de la matrice par sa transposée ou matrice des cosinus entre vecteurs-lignes). A l'issue des transformations, une nouvelle table de hachage du corpus est créée avec les mêmes clés et traits sémantiques, ainsi que les nouvelles distributions.

Dans un deuxième temps, le mot de référence est recherché dans la table de hachage du corpus. Pour ce faire, le programme utilise les traits sémantiques du mot de référence étudié et les recherche dans la table du corpus.
Il récupère ensuite dans la table du corpus les indices de lignes des traits identifiés. A partir de cette liste d'indices, le programme extrait les colonnes correspondantes de la matrice et obtient ainsi la matrice corpus-mot. Il effectue ensuite une opération similaire à partir du cotexte : utilisation des traits sémantiques véhiculés par les mots du cotexte du mot de référence. Mais cette fois, le programme sélectionne les lignes et non les colonnes de la matrice corpus-mot.

Il exporte enfin les données de la matrice obtenue dans un fichier au format txt.

Ajoutons que les matrices et tables de hachage sont, à chaque transformation, stockées sous forme sérialisée, c'est-à-dire dans des fichiers contenant l'ensemble des informations permettant de les reconstituer.

5.1.2) Justification des choix effectués

Quelques explications sur les choix de programmation s'imposent.

J'ai hésité sur le langage de programmation. Deux alternatives s'offraient : programmation en Java ou programmation en Python. Rappelons que le programme Sémy est écrit en Python, donc que ce langage aurait permis d'intégrer les classes de Sémy à celles de mon programme. Cependant, je n'avais aucune base en Python. L'initiation à un nouveau langage aurait certes pu être très formatrice, mais risquait de s'effectuer au détriment de l'efficacité. J'ai donc préféré Java et j'ai ainsi pu y approfondir mes connaissances. La communication entre Python et Java reste au demeurant toujours réalisable, grâce à Jython, un interpréteur de Python écrit en Java.

Le nombre de classes dont je dispose est réduit. En effet, une multiplication des classes disponibles implique une multiplication des objets créés et stockés en mémoire. Or le manque d'espace mémoire a été un problème crucial, d'où le choix effectué.

La *classe Matrice* a été bâtie à partir d'un package Java de classes matricielles, disponible à l'adresse http://math.nist.gov/javanumerics/jama/. Ce package permet de créer un objet Matrix à partir d'un tableau de valeurs et d'effectuer les opérations de base (addition, soustraction,…) ainsi que certaines opérations plus compliquées, comme la décomposition en valeurs singulières, indispensable dans LSA. Ma classe Matrice reprend un certain nombre des fonctionnalités du package Jama et enrichit celui-ci de méthodes spécifiques aux transformations que je souhaite effectuer : opérations matricielles permettant de réaliser tf-idf, LSA, de calculer des cosinus entre vecteurs-lignes, la matrice des cooccurrences ou les valeurs théoriques du test du χ^2. A l'origine, ma classe Matrice héritait donc de la classe Matrix du package Jama, mais pour des raisons de mémoire (création d'objets Matrix intermédiaires consommateurs d'espace), j'ai dupliqué les éléments de Matrix qui paraissaient utiles dans Matrice, devenue indépendante du package Jama.

Les tables de hachage ont été introduites pour faire le pont entre les lignes de la matrice et les traits sémantiques auxquels elles correspondent. En effet, pour manipuler les objets de type Matrice, il faut disposer d'une indexation chiffrée des lignes, alors que l'analyse des résultats impose d'identifier les traits sémantiques correspondant à chaque ligne.

La sérialisation des matrices ou tables de hachage présente un double avantage. D'une part, les paramètres à faire varier et donc les combinaisons des opérations à réaliser sont nombreux. Or certaines opérations sont coûteuses en temps. La sérialisation d'objets intermédiaires permet ensuite de repartir du stade intermédiaire correspondant pour faire en suite varier les paramètres souhaités. Par exemple, l'extraction de l'image de n cotextes différents de la matrice de cooccurrences pourra s'opérer à partir de cette matrice sans qu'il soit nécessaire d'effectuer toutes les étapes préalables à chaque fois. Ajoutons que la sérialisation constitue une forme de sauvegarde des résultats et évite de perdre tous les résultats antérieurs en cas de problèmes dans les dernières étapes. D'autre part, cette sérialisation évite la multiplication des objets stockés en mémoire et permet donc de remédier à certains problèmes d'espace mémoire. Par exemple, supposons que l'on souhaite générer trois matrices de cooccurrences, une à partir d'une transformation tf-idf, une à partir d'une transformation de type LSA et une sans aucune transformation, puis que l'on veuille sélectionner dans chacune de ces matrices la représentation du mot et de son cotexte. L'espace mémoire occupé est à peu près le triple de celui consommé par une seule matrice. La sérialisation, mise à zéro temporaire puis désérialisation des matrices non utilisées permet de gérer la surconsommation d'espace mémoire. Soulignons toutefois que la sérialisation et désérialisation des objets adaptés pour chaque transformation impliquent plus de manipulations et compliquent donc la gestion du programme. Cette solution est acceptable dans un premier temps, phase expérimentale où prime la génération de données, mais

nécessite d'être révisée pour permettre à n'importe quel utilisateur de se servir facilement du programme.

Enfin, la création de fichiers de données au format texte ou csv en sortie ouvre la porte à l'exploitation des résultats par d'autres logiciels, qui souvent prennent en entrée des fichiers à ces formats-là. Ce sera par exemple le cas du logiciel PermutMatrix utilisé pour l'analyse des données et décrit au paragraphe 5.3.1.

5.1.3) Limites et difficultés rencontrées

Le principal problème rencontré est celui de la mémoire. En effet, dès que le nombre de traits sémantiques et nombre de sous-unités du corpus devient conséquent, les matrices et tables de hachages deviennent des objets particulièrement volumineux au niveau de la mémoire. Ainsi, pour un corpus contenant plus de 8000 traits sémantiques, le fichier de la matrice sérialisée des cooccurrences (matrice de dimension supérieure à 8000 sur 8000) ou celui de sa table de hachage fait une taille de plus d'un demi Gigaoctet. Sur des machines disposant de 2 Gigaoctets de RAM, le seuil de capacité est rapidement atteint avec quelques objets de ce type.
Ce problème de mémoire a nécessité des adaptations du programme informatique (pas d'héritage de la classe Matrix, sérialisations, fragmentation du contenu du main, …). Malgré ces adaptations, le programme n'a pu traiter des corpus au-delà d'une certaine taille. Deux critères déterminent le seuil limite : le nombre de traits sémantiques et le nombre d'unités de découpage du corpus d'origine. Quelques essais ont permis d'établir un seuil entre 8850 et 9000 traits sémantiques pour 628 unités de découpage du corpus. Soulignons enfin que le seuil déterminé ne prend pas en compte un aspect : l'exportation des fichiers texte de la matrice de cooccurrences. Plus précisément, la matrice de cooccurrences peut être calculée et sérialisée, donc il est possible d'avoir accès à ses informations. En revanche, le fichier texte qui permet de visualiser l'ensemble des résultats est trop important pour être généré. Ce problème n'est pas fondamental puisque d'une part la matrice de cooccurrences n'est jamais exploitée directement mais à travers des sous-matrices, d'autre part l'exportation de sous-matrices de la matrice de cooccurrences fonctionne et pourrait permettre, moyennant quelques lignes de code supplémentaires, de reconstituer la matrice d'origine si l'accès à celle-ci devait s'avérer nécessaire.

Autre problème à mentionner : la lenteur du programme, problème croissant avec la taille du corpus. Pour y remédier, une phase d'optimisation a été nécessaire, avec par exemple reprise des boucles *for*, de l'emplacement où les objets étaient créés et surtout changement au niveau du traitement des entrées et sorties. Le changement principal repose sur l'introduction d'éléments bufferisés pour l'écriture (objet StringBuilder) et la lecture de fichiers. La rapidité du programme a ainsi pu être fortement accrue.

5.2) Paramètres des tests effectués

5.2.1) Les supports de référence

Les tests effectués se sont appuyés sur un corpus constitué de six contes extraits de Wikisource :

- George Sand, *La fée poussière* : une fée, apparemment pauvre petite vieille et dans ce conte allégorie de la poussière, emmène la narratrice dans son palais et lui fait voir ses richesses ainsi que les rouages de création du monde qu'elle y met en œuvre
- Charles Renel, *La Race inconnue – l'enfant d'argile* : conte africain sur une femme stérile qui fait croire à son village qu'elle a accouché de jumeaux, dont un est fils d'un esprit protecteur
- Ernest du Laurens de la Barre, *Fantômes bretons – les poires d'or* : une famille possède un poirier produisant des poires en or qui, une fois prêtes à être cueillies, se font toujours voler, elle découvre que le voleur est un ogre ; un des fils de la famille part réclamer les poires volées auprès de l'ogre, finit par tuer celui-ci et épouser sa fille

- Hans Christian Andersen, *Une rose de la tombe d'Homère* : conte sur l'amour malheureux d'un rossignol pour une rose et leur destin tragique
- Madame d'Aulnoy, *Le nain jaune :* un vil nain jaune se fait promettre en épouse une belle princesse qui trahit la promesse, se donne en mariage à un roi et se fait enlever par le nain
- Jacob et Wilhelm Grimm, *Hänsel et Gretel* : deux enfants pauvres, un frère et une sœur, sont abandonnés par leurs parents dans la forêt, aboutissent chez une sorcière qui veut dévorer le frère, parviennent à éliminer celle-ci et retourner chez eux chargés de richesses.

Ces contes ont été rédigés par six auteurs différents. Ils reprennent un certain nombre d'éléments traditionnels des contes, comme les personnages royaux (belle princesse, reine, roi), les créatures merveilleuses (nain, géant, fée, sorcière), les palais et châteaux, le thème de la magie, de la richesse, de la pauvreté, de l'amour, de la bravoure ou encore de la fécondité. Mais ils balayent un vaste champ, varient les approches et comportent une richesse thématique et lexicale suffisante. Le découpage d'origine, en paragraphes, a été conservé comme découpage de référence au cours des expérimentations. Le corpus comporte 240 unités de découpage, 20324 occurrences de mots (approximativement 28 pages en taille 12, police Times New Roman dans un éditeur de textes) et 8467 traits sémantiques différents (ou famille de traits sémantiques).

La liste suivante comporte les mots de référence étudiés. Le choix des mots s'est porté sur des noms concrets, au sémème diversifié, c'est-à-dire composé d'unités sémantiques très différentes et susceptibles d'être activées ou inhibées selon des cotextes particuliers.

- pollen, de taille 18 (taille du sémème en nombre de familles de traits sémantiques, après traitement effectué par Sémy)
- nacre, de taille 29
- sable, de taille 34
- éclat, de taille 46
- fer, de taille 74
- or, de taille 91
- rose, de taille 103 (fleur et couleur regroupés)

A chaque mot ont été associés un à quatre cotextes, centrés sur les mots et sélectionnés manuellement. Les cotextes correspondent à des paragraphes ou regroupements de phrases. La taille des cotextes varie de 500 à 1654 traits sémantiques. Ces cotextes sont disponibles en annexe 3.

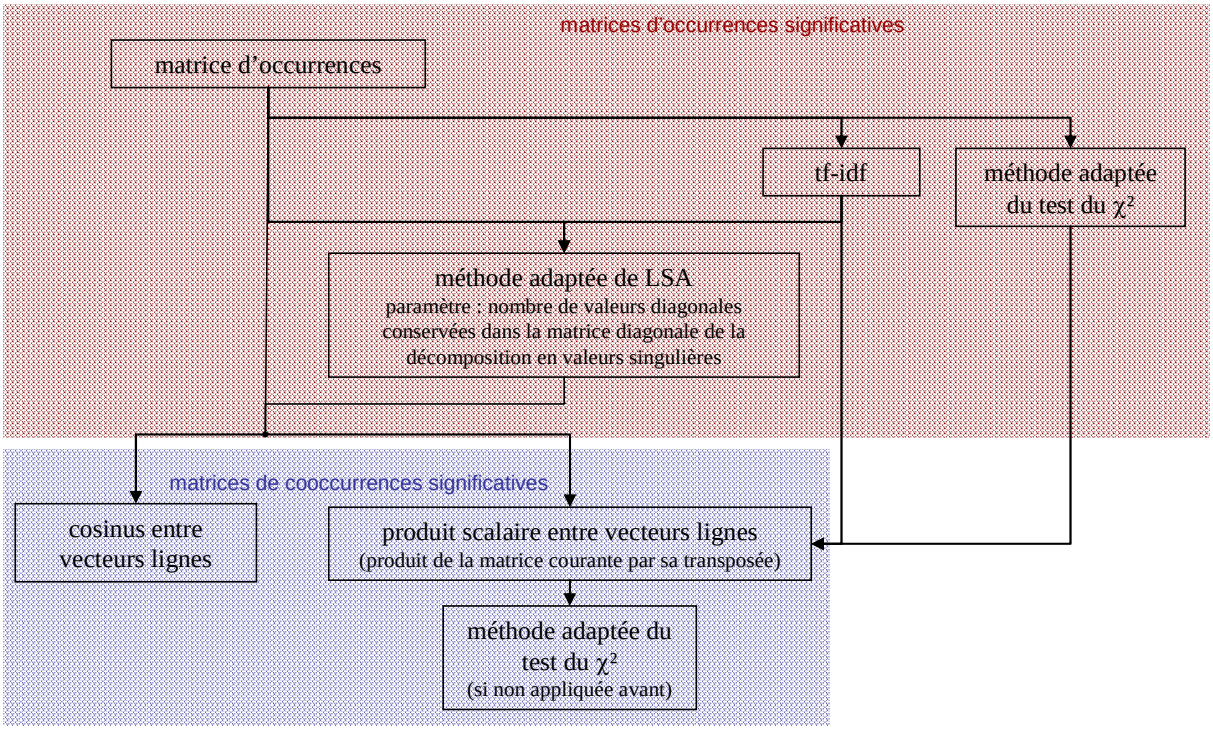
Un autre corpus a été généré à partir de la base de textes Frantext. Il est constitué d'extraits de romans de quatre auteurs du XIXe et XXe siècle :

- Marcel Proust, *A la recherche du temps perdu – Sodome et Gomhorre*
- Maurice Genevoix, *La boîte à pêche*
- Jean Giraudoux, *Suzanne et le Pacifique*
- Maurice Schreiber, *Un silence d'environ une demi-heure*

Ce corpus est centré sur pollen, avec 26 cotextes d'apparition de ce mot. L'exploitation de ce corpus n'a pu être menée à bien pour des raisons d'espace mémoire : le corpus contient plus de 12000 traits, ce qui excède largement le seuil limite. Un travail d'élagage des textes permettrait d'obtenir une taille adaptée. La priorité a cependant été donnée au corpus de contes, dont les pistes d'exploitation se sont avérées particulièrement riches et prioritaires sur des études comparatives entre les deux corpus mentionnés ou de nouvelles études exclusivement sur le corpus de Frantext.

5.2.2) Opérations mathématiques appliquées

A mot et cotexte fixé, différentes transformations mathématiques ont été appliquées, afin de comparer leurs effets et déterminer lesquelles seraient les plus appropriées selon les observations à faire (activation et inhibition de traits sémantiques, structuration du sémème). Le schéma ci-dessous retranscrit les combinaisons de transformations effectuées :



La méthode adaptée du χ^2 reprend les calculs mis en œuvre par [Victorri, 2005, p. 119], à un changement près : au lieu d'appliquer une fonction linéaire par morceau au rapport $\frac{m_{ij}}{n_{ij}}$ du rapport de la valeur théorique sur la valeur moyenne, je l'ai appliquée au rapport inverse $\frac{n_{ij}}{m_{ij}}$. En effet, la formule proposée ne paraissait pas cohérente avec le rôle de la fonction ni avec le tableau de résultats [Victorri, 2005, p. 120 Tableau 3.2]. Celle-ci doit mesurer un degré d'affinité entre deux mots, d'autant plus important que la valeur réelle est supérieure à la valeur théorique. Par mesure de précaution, j'ai également effectué quelques expériences avec la fonction appliquée au rapport $\frac{m_{ij}}{n_{ij}}$.

La méthode adaptée du χ^2 a été appliquée à deux niveaux différents : dans un cas avant le calcul de la matrice de cooccurrences, dans le second cas après celle-ci. Les perspectives sont différentes dans chaque cas. Lorsque la méthode est appliquée avant calcul de la matrice de cooccurrences, on effectue un filtrage par rapport à la répartition des traits sémantiques par paragraphe. Un trait sémantique sera considéré comme significatif, et affecté d'un coefficient reflétant cette significativité, s'il est surreprésenté par rapport à une distribution équiprobable par paragraphe. La constitution de la matrice de cooccurrences revient à considérer que les traits sémantiques qui sont surreprésentés ou sous-représentés dans les mêmes paragraphes ont un fort degré d'affinité. L'autre cas de figure reprend la même approche que [Mauceri, 2007] puisqu'il effectue un filtrage sur la matrice de cooccurrences, en aval de la chaîne de transformations. Un degré d'affinité entre deux traits sémantiques est significatif s'il est supérieur à un degré d'affinité théorique. Celui-ci est obtenu sous hypothèse d'équirépartition des affinités des traits ou encore d'indépendance des traits sémantiques en termes d'affinité.

La méthode adaptée de LSA reprend la partie centrale des opérations effectuées dans LSA, à savoir la décomposition en valeurs singulières et la mise à 0 d'un certain nombre de coefficients diagonaux de la matrice diagonale obtenue lors de la décomposition. Cette matrice diagonale est, dans l'expérience effectuée sur les contes, de taille 240. Les expériences ont été menées pour 5, 10, 25 et 50 valeurs diagonales conservées à leur valeur d'origine, les autres étant annulées. La combinaison de tf-idf et de la méthode adaptée de LSA permet de se rapprocher du véritable modèle utilisé par LSA.

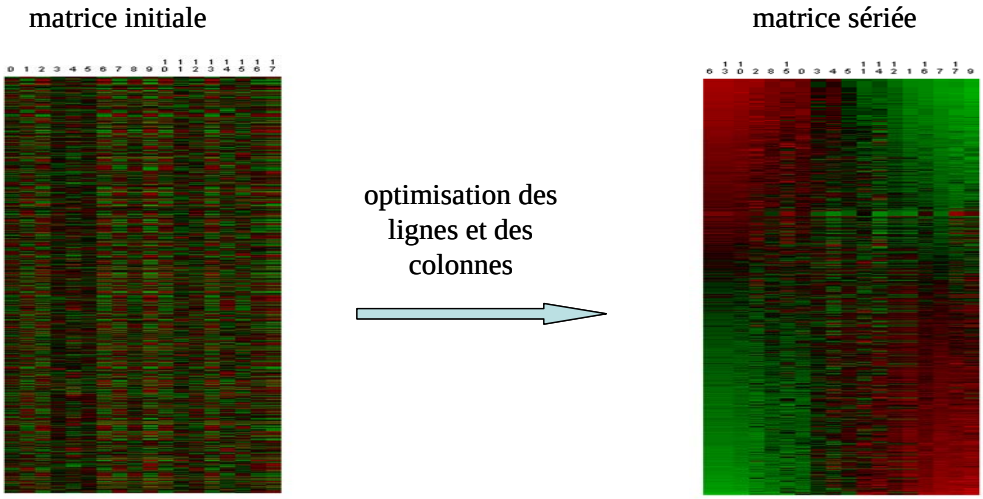
5.3) Tests et analyse des résultats

5.3.1) Méthodes d'analyse mathématiques

5.3.1.1) Visualisation des matrices : logiciel PermutMatrix

PermutMatrix est un logiciel développé par [Caraux & Pinloche, 2005] dans le cadre de la bioinformatique. Il permet de visualiser et analyser de matrices. Il procède par réagencement des lignes et colonnes de la matrice et s'appuie sur différentes méthodes : des méthodes de classification hiérarchique et des méthodes de sériation. J'ai utilisé ce logiciel pour observer les sèmes du mot activés et inhibés, j'ai donc privilégié les méthodes de sériation

Les méthodes de sériation partent de l'hypothèse qu'il existe un ordre qu'elles essayent de faire émerger. Elles reposent sur un critère à optimiser. J'ai choisi la méthode dite de positionnement unidimensionnel pour mes applications, efficace lorsqu'il existe une structure 1D. Je cherchais à établir les sèmes dominants et inhibés, c'est-à-dire une relation du type « la significativité du trait 1 est supérieure à celle du trait 4, elle-même supérieure à celle du trait 2,… », donc à déterminer une structure unidimensionnelle entre les traits sémantiques, d'où le choix de cette méthode. Sur le plan mathématique, elle part des coefficients d'une matrice D de dissimilarité. Celle-ci contient en position (i,j) la distance (euclidienne dans mon cas) entre la ligne (resp. colonne) i et la ligne (resp. colonne) j dans le cas d'une optimisation selon les lignes (resp. colonnes). Le critère à minimiser est : $C(\pi) = \sum\sum\left(d\left(\pi(i),\pi(i)\right) - a\left|i-j\right|\right)^2$, où π est une permutation et a une constante multiplicative de mise à l'échelle.



5.3.1.2) Analyse de moyennes et écarts-types

J'ai par ailleurs complété mes analyses par divers calculs de moyennes et écarts-types, lorsque les résultats de PermutMatrix ne permettaient pas une interprétation immédiate ou lorsque certaines conclusions tirées méritaient d'être renforcées par des observations complémentaires.

5.3.2) Tests réalisés : observations des activations et inhibitions

Les résultats générés par le programme Java ouvraient la porte à de nombreuses expériences et de multiples axes d'observation. Les efforts se sont portés sur un petit ensemble d'aspects seulement et ne recouvrent qu'une petite partie des pistes à explorer décrites dans ce rapport.

Une première tentative a été effectuée sur le corpus de Frantext, construit autour du pollen. Vingt-six contextes ont été sélectionnés et annotés : trois personnes différentes ont déterminé quels traits du sémème de pollen étaient activés selon les différents contextes. Cette analyse humaine devait valider les résultats de l'expérience. Celle-ci n'a pu être menée à bien car le corpus, de taille trop importante, a révélé les problèmes de mémoire, d'où la nécessité de construire le petit corpus de contes. Cependant, la démarche entreprise sur Frantext a été imitée sur le corpus de contes, pour 13 contextes : le repérage des traits sémantiques activés dans chaque contexte a fait l'objet d'une analyse humaine préalable. Cette démarche a été entreprise pour éviter une démarche inverse à la validation des résultats par l'analyse humaine : il s'agissait de comparer l'influence des différents paramètres et d'étudier les failles du modèle, et non pas de tirer des conclusions sur les phénomènes linguistiques à partir des résultats mathématiques, démarche qui ne peut être effectuée qu'après la première, après établissement d'un modèle valide sur de petites expériences.

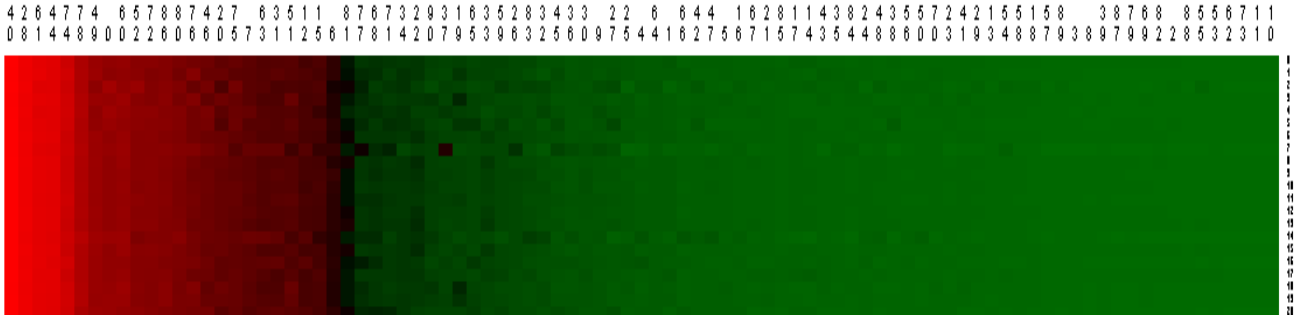
Les premières analyses portent sur l'influence des différentes transformations mathématiques. Dans un second temps, j'ai étudié l'influence des contextes, sous l'hypothèse que la matrice contexte-mot contenait des informations dues à la matrice corpus – mot, mais que l'influence locale était suffisante pour faire émerger des différences significatives d'un contexte à l'autre. Les observations faites ont remis en cause cette hypothèse et conditionné les expériences suivantes, constituées d'abord de la recherche de facteurs explicatifs, puis d'une méthode pour observer les variations fines.

5.3.2.1) Analyse n°1 : influence de la transformation mathématique

La première série d'analyses a pour but de faire émerger les effets liés à chaque transformation mathématique, à mot et cotexte fixés. Nous avons d'abord étudié l'activation et l'inhibition des traits sémantiques. Les observations faites portaient sur les traits sémantiques d'éclat dans le cotexte n°10 (voir annexe 3), or dans le cotexte n°10, sable dans le cotexte n°4 et pollen dans le contexte n°5.

Cooccurrences simples, sans autre transformation

Dans tous les exemples analysés, les familles de traits sémantiques activés se regroupent en deux catégories. La première catégorie est constituée de familles morphologiques non pertinentes, d'une taille démesurée due d'une part au regroupement de familles sémantiquement distinctes qui auraient dû être réparties dans plusieurs classes, d'autre part à la généralité du sens porté par les représentants de chaque famille. La seconde catégorie est constituée de traits sémantiquement faibles, de sens très général, comme /objet/ ou /faire/. A titre d'illustration, voici la matrice de cooccurrences simples du mot or obtenue par PermutMatrix après réorganisation optimale des colonnes :



Indice des 5 traits activés (5 premières colonnes)	Représentant de la famille morphologique	Taille de la famille morphologique
40	représentant impossible à identifier (trop de diversité)	859 éléments
28	voir, -vis-	-
61	prendre, entrer, produire, ouvrir	144 éléments

44	faire	-
74	forme, fond	182 éléments

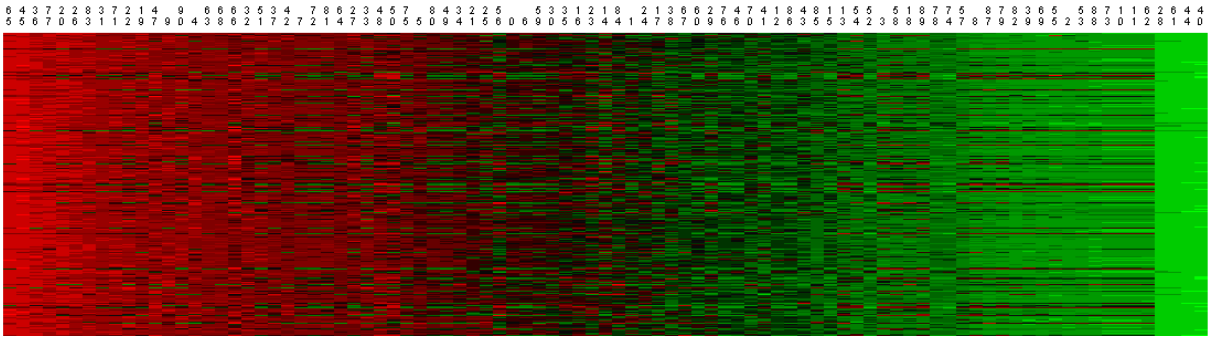
Les traits inhibés correspondent à des termes rares en langue française, comme /anthère/ et /utricule/, traits sémantiques de pollen les plus inhibés de son sémème, ou à des traits inattendus dans un corpus de contes, mais dont l'apparition pourrait se justifier dans d'autres corpus. Par exemple, les traits sémantiques /acide/, /ductile/, /atome/ ou /nickel/ renvoient aux propriétés physiques de l'or, renvoient au domaine scientifique et font partie des sèmes les plus inhibés du mot or.

L'absence de transformation donne donc des résultats satisfaisants au niveau des traits inhibés. En revanche, il fait apparaître comme dominants des traits qu'on pourrait qualifier de traits sémantiquement faibles et qui devraient être au contraire inhibés.

Méthode tf-id

La méthode tf-idf permet de restructurer le sémème et surtout les traits dominants, point faible du calcul des cooccurrences simples. Une grande partie des traits activés sont des traits sémantiquement forts, au contenu sémantique riche. Soulignons que certains traits ne répondent pas à ces caractéristiques et sont des traits généraux, comme /très/ dans le sémème de pollen.

Les sèmes inhibés sont à peu près les mêmes que ceux obtenus sans transformation particulière, auxquels s'ajoutent les familles dominantes de la matrice des cooccurrences simples, à savoir les familles trop grandes ou représentatives de « traits-outils », c'est-à-dire de termes extrêmement généraux. La matrice obtenue par PermutMatrix sur le sémème d'or est un exemple type des observations décrites :



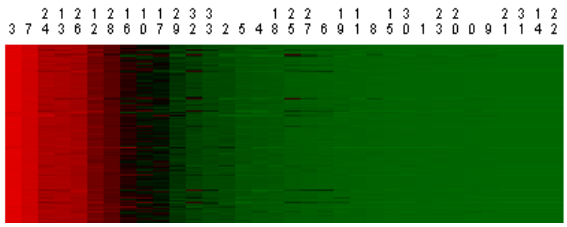
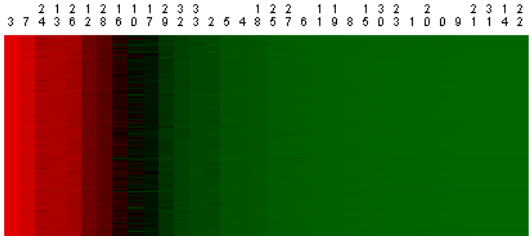
Indice des 10 traits dominants (10 premières colonnes)	Représentant de la famille morphologique	Indice des 10 traits les plus inhibés (de la droite vers la gauche)	Représentant de la famille morphologique
65	eau	40	représentant impossible à identifier (trop de diversité)
45	lumière	44	faire
36	métal	61	prendre, entrer, produire, ouvrir
20	représenter	28	voir, vis-
77	couleur	62	galon
26	chaud	11	inaltérable
83	argent	10	pépite

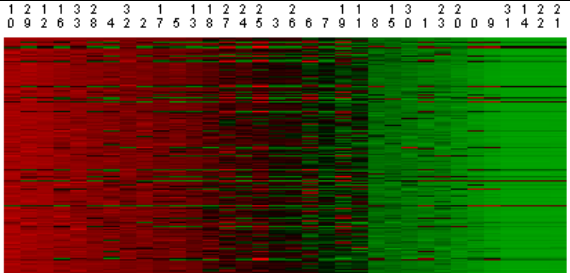
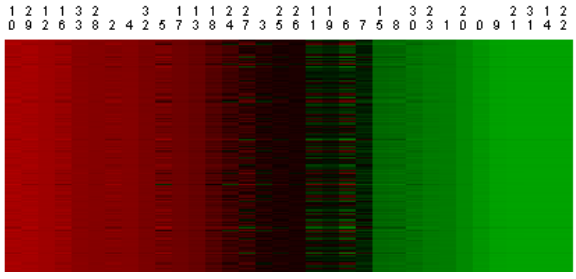
22	allié / alliance / alliage	73	nickel
71	civil / civiliser	88	acide
31	utile / utiliser	53	étalon

Les résultats obtenus par la méthode tf-idf sont donc beaucoup plus probants que ceux obtenus sans transformation. De plus, la méthode répond aux attentes : elle inhibe les traits surreprésentés tout en conservant un faible poids pour les termes rares. La transposition de la méthode tf-idf des mots aux traits sémantiques apparaît donc comme valide et efficace.

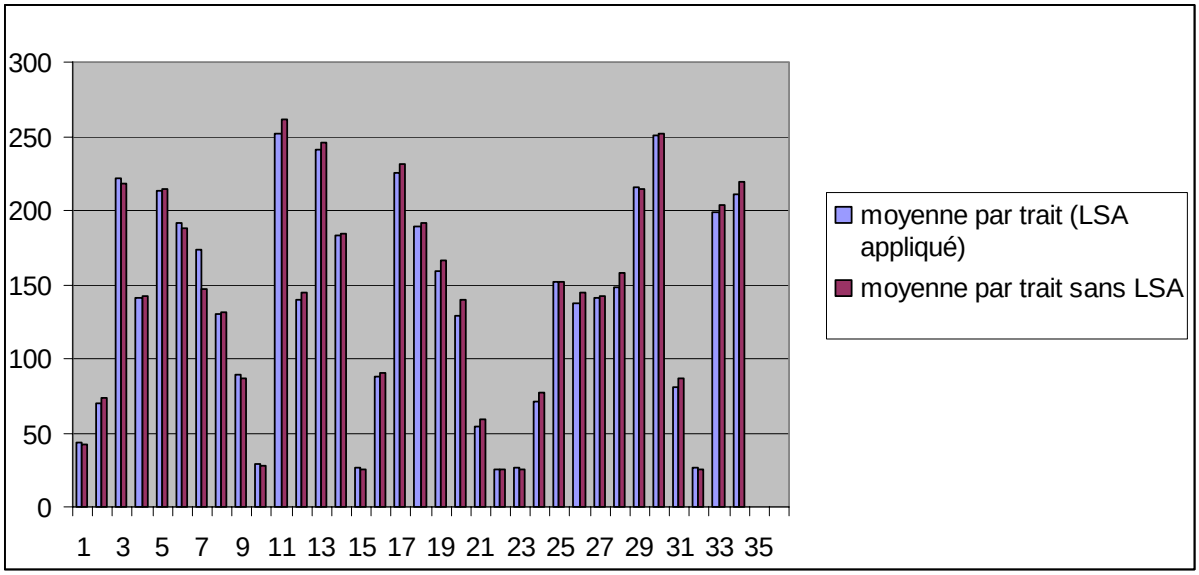
Méthode adaptée de LSA

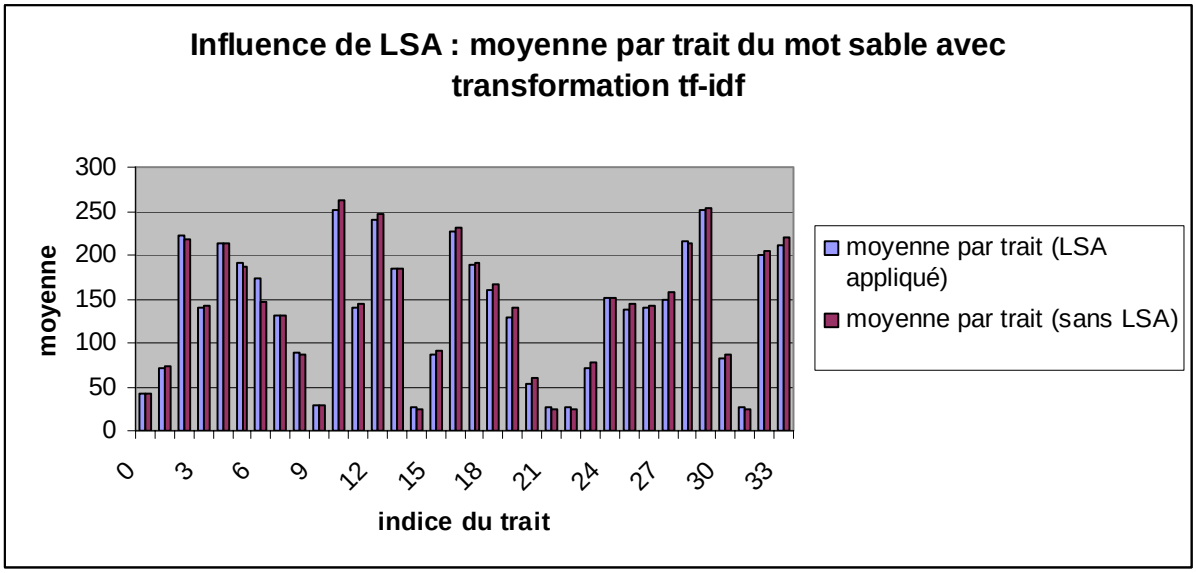
La méthode de type LSA (partie centrale de LSA, avec décomposition en valeurs singulières et projection sur les directions principales de la matrice diagonale) appliquée seule avant le calcul des cooccurrences, respectivement après application de tf-idf, n'a pas donné de résultats notablement différents de ceux des cooccurrences simples (resp. de la méthode tf-idf). En effet, l'allure générale de la matrice finale obtenue par PermutMatrix est la même avec ou sans la méthode de type LSA, ainsi que la liste des sèmes activés, inhibés et l'ordre des colonnes les représentant. Signalons l'une ou l'autre permutation mineure sur des paires de colonnes consécutives, quelquefois des triplets de colonnes consécutives, mais rarement plus. Les matrices ci-dessous correspondent au sémème de sable dans le contexte 3.

Calcul des cooccurrences sans LSA	Calcul des cooccurrences avec LSA
	

Application de tf-idf puis calcul des cooccurrences sans LSA	Application de tf-idf puis calcul des cooccurrences avec LSA
	

L'observation des moyennes des coefficients pour chaque trait sémantique rejoint les constations précédentes : les moyennes ne sont pas significativement différentes. L'histogramme ci-dessous repésente les moyennes des coefficients de la matrice contexte-mot pour sable dans le contexte 4. L'application de LSA fait suite à celle de tf-idf.





La méthode adaptée de LSA fait donc apparaître des variations fines, mais insuffisantes pour changer en profondeur la structure du sémème. Ces résultats ne sont cependant pas anormaux. Alors que dans la méthode tf-idf, significativité et fréquence avaient des sens de variation inverses au-delà d'un certain seuil, LSA a en quelque sorte pour vocation de lisser les variations observées, non de les restructurer complètement. Pour peu que le nuage de points soit allongé le long des directions principales, les changements apportés par LSA seront minimes. L'étude des coefficients de la matrice diagonale obtenue lors de la décomposition en valeurs singulières permettrait une analyse plus fine des résultats.

Méthode adaptée du χ^2 (appliquée à la matrice de cooccurrences)

Dans l'exemple analysé (éclat dans le cotexte n°10), la méthode adaptée du χ^2 donne des résultats satisfaisants concernant les familles de traits à faible contenu sémantique ou les familles de traits de taille excessive, puisque ces familles se trouvent inhibées, comme la famille de /donner/, /faire/ ou encore l'adverbe /surtout/. En revanche, les traits sémantiques rares, comme /touffe/ ou /tapage/ voient leur significativité relative remonter par rapport à toutes les autres transformations et se situent dans des positions centrales à dominantes. Ainsi, les quelques observations sur la transformation adaptée du χ^2 laissent entrevoir une capacité à inhiber les traits non pertinents à contenu sémantique trop général, mais l'affectation d'une trop forte significativité aux termes rares.

Calcul des cosinus

Tout comme pour la matrice de cooccurrences obtenue en multipliant la matrice d'occurrences par sa transposée, le calcul des cosinus fait émerger comme traits dominants les familles de traits sémantiques de taille particulièrement importante ou les traits à faible contenu sémantique (par exemple, /état/ ou /caractère/). Le calcul des cosinus semble même encore moins adapté que le calcul des cooccurrences simples pour faire émerger les traits dominants et inhibés car certains traits à forte significativité sur le plan linguistique semblent plus inhibés par le calcul des cosinus. Ce phénomène s'explique par le critère de calcul des coefficients de la méthode cosinus : seul compte l'angle fait entre deux vecteurs – lignes de traits sémantiques, quel que soit la norme de ces vecteurs. On perd totalement l'information sur la quantité d'occurrences absolue, seule compte la similarité de distribution des traits sur les différents paragraphes. Au vu des résultats, on serait donc plutôt incité à prendre le produit scalaire plutôt que le cosinus entre vecteurs-lignes pour calculer des coefficients de similarité, du moins sans transformation préalable. L'étude de l'influence du cosinus après tf-idf suivie de la méthode adaptée de LSA témoigne également en défaveur du calcul du cosinus, avec une remontée brutale de termes non pertinents auxquels tf-idf avait affecté une faible significativité. Par exemple, sur le mot **éclat** dans le contexte 10, considérons les deux familles suivantes, extraites du sémème :

- famille de /faire/ : faillibilité,NOM refaçonnement,NOM défaitisme,NOM redéfaire,VERBE défaitiste,ADJ fabrique,NOM affaireux,ADJ fabricatrice,NOM préfabriqué,ADJ méfaire,VERBE refait,NOM fabrication,NOM refaiseuse,NOM défaitiste,NOM factieuse,NOM méfait,NOM faillir,VERBE défaillance,NOM façonnerie,NOM facturer,VERBE fautivement,ADV défaillement,NOM faillite,NOM façonnier,NOM défaçonner,VERBE falloir,VERBE défaillir,VERBE fabricant,NOM affaire,NOM refaiseur,NOM faillie,NOM préfabriqué,NOM affairiste,NOM fabricateur,NOM défait,ADJ fait,NOM faillible,ADJ factionnaire,NOM facturation,NOM préfabrication,NOM refaire,VERBE réfection,NOM refaçonner,VERBE préfabriquer,VERBE fauter,VERBE parfaire,VERBE surfacturer,VERBE refaçonnage,NOM facture,NOM faction,NOM défaillant,ADJ fabriquer,VERBE faute,NOM failli,NOM façonnier,ADJ façon,NOM faire,NOM refabriquer,VERBE factieux,NOM façonnage,NOM faire,VERBE malfaçonné,ADJ défaite,NOM fabricante,NOM façonnière,NOM fabricien,NOM réfectionner,VERBE façonner,VERBE failli,ADJ fautif,ADJ factionnaire,ADJ défaire,VERBE refabrication,NOM défaut,NOM refaçonneur,NOM surfacturation,NOM factieux,ADJ façonnement,NOM
- famille de /prendre/, /ouvrir/, /poser/ et /produire/ (regroupement non pertinent car fusion de plusieurs familles) : /1101/ : preneur,NOM entr'ouvrir,VERBE reproductibilité,NOM décomposant,ADJ reprocheur,ADJ productif,ADJ composant,ADJ représenté,ADJ produire,VERBE improduit,ADJ incompréhensiblement,ADV entrouvrir,VERBE reproductivité,NOM mécomprendre,VERBE production,NOM rentré,NOM entrance,NOM représentation,NOM surprise,NOM appréhendement,NOM emprisonné,ADJ autoreproducteur,ADJ rentrayeur,NOM rentrant,NOM reprisage,NOM mécompréhension,NOM reproductif,ADJ incompréhensible,ADJ apprésentation,NOM prisonnier,NOM compréhension,NOM entr'ouvrement,NOM reproduire,VERBE plexus,NOM reproductrice,NOM pris,ADJ sentimentaliste,NOM sentimentalité,NOM surproduction,NOM entrouverture,NOM entreprise,NOM indécomposé,ADJ déprise,NOM reprographique,ADJ senti,NOM incomplexe,ADJ sentimentalisation,NOM irreprésentable,ADJ rentrante,NOM incompréhensif,ADJ reproche,NOM prisonnière,NOM rentrayeuse,NOM présent,ADJ prison,NOM représentée,NOM prise,NOM représenter,VERBE présenter,VERBE repriser,VERBE incompris,ADJ décomposer,VERBE reprise,NOM reprocher,VERBE présence,NOM procès,NOM compréhensible,ADJ représentante,NOM preneuse,NOM reproductivement,ADV entrer,VERBE reprisable,ADJ rentrant,ADJ représentativité,NOM reprographier,VERBE prendre,VERBE rentrage,NOM indécomposable,ADJ coproduction,NOM déprendre,VERBE repriseur,ADJ reprochable,ADJ reprendre,VERBE imprenable,ADJ

composante,NOM présentation,NOM reproductible,ADJ improductivement,FUNC repriseuse,NOM coproduire,VERBE improductif,ADJ preneur,ADJ producteur,NOM répréhension,NOM préhension,NOM représentable,ADJ senti,ADJ entrant,NOM comprendre,VERBE reproduction,NOM rentrure,NOM entreprendre,VERBE incompréhensibilité,NOM appréhension,NOM sentimentaliser,VERBE produit,ADJ surproduit,NOM improductivité,NOM sentiment,NOM complexe,ADJ représenté,NOM représentatif,ADJ sentimental,ADJ prisonnier,ADJ produit,NOM reproducteur,ADJ intercompréhension,NOM sentimentalisme,NOM entrée,NOM rentrée,NOM sentir,VERBE emprisonner,VERBE dissentiment,NOM surproduire,VERBE appréhender,VERBE rentrer,VERBE rentré,ADJ surreprésentation,NOM apprésenter,VERBE entrant,ADJ présent,NOM composant,NOM représentativement,FUNC reproducteur,NOM incompréhension,NOM décomposition,NOM entrante,NOM entrepreneur,NOM emprisonnement,NOM surprendre,VERBE appréhendé,ADJ représentant,NOM reprographie,NOM décomposable,ADJ

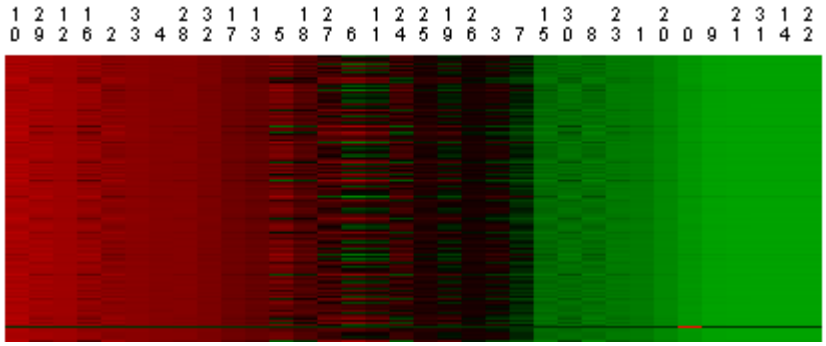
Ces familles sont précisément celles auxquelles on souhaite affecter un faible coefficient de significativité. Dans l'application de tf-idf suivie de LSA et calcul de cooccurrences par produit scalaire, ces deux familles sont inhibées, respectivement en 36^{e} et 39^{e} positions (classement des colonnes rouges, traits à forte significativité, vers les colonnes vertes, traits à faible significativité) pour un sémème de 46 familles de traits. En revanche, lorsque le produit scalaire est remplacé par un calcul de cosinus, ces familles de traits se retrouvent respectivement en 5^{e} et 1ère positions.
Réitérer les expériences, relativement peu nombreuses sur le cosinus, permettrait d'obtenir des résultats plus fiables. Cependant, les analyses réalisées donnent à penser que le cosinus n'est pas la méthode la plus pertinente pour faire émerger la significativité.

5.3.2.2) Influence des contextes

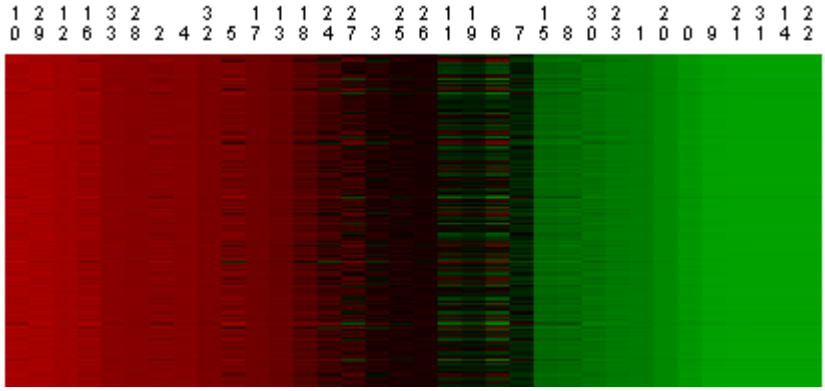
Analyse n°1 : comparaison des contextes par PermutMatrix et indicateurs de valeurs centrales et dispersion

Les indicateurs utilisés, à savoir PermutMatrix ou des calculs d'indicateurs moyens et de dispersion, font apparaître quelques différences entre contextes mais très faibles, quelle que soit la transformation. En effet, les mêmes plages de couleurs sont affectées aux mêmes séries d'indices de traits sémantiques. On peut observer des permutations entre rangs successifs, essentiellement sur 3 à 4 rangs et un peu plus dans quelques cas rares. Ainsi, l'agencement des traits reste dans son ensemble à peu près identique. L'exemple ci-dessous correspond au mot sable dans les contextes 2,3 et 4 pour la transformation tf-idf suivie de LSA :

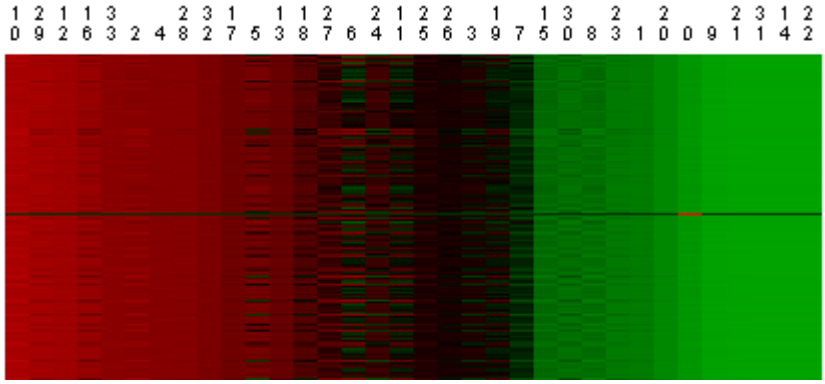
Contexte 2



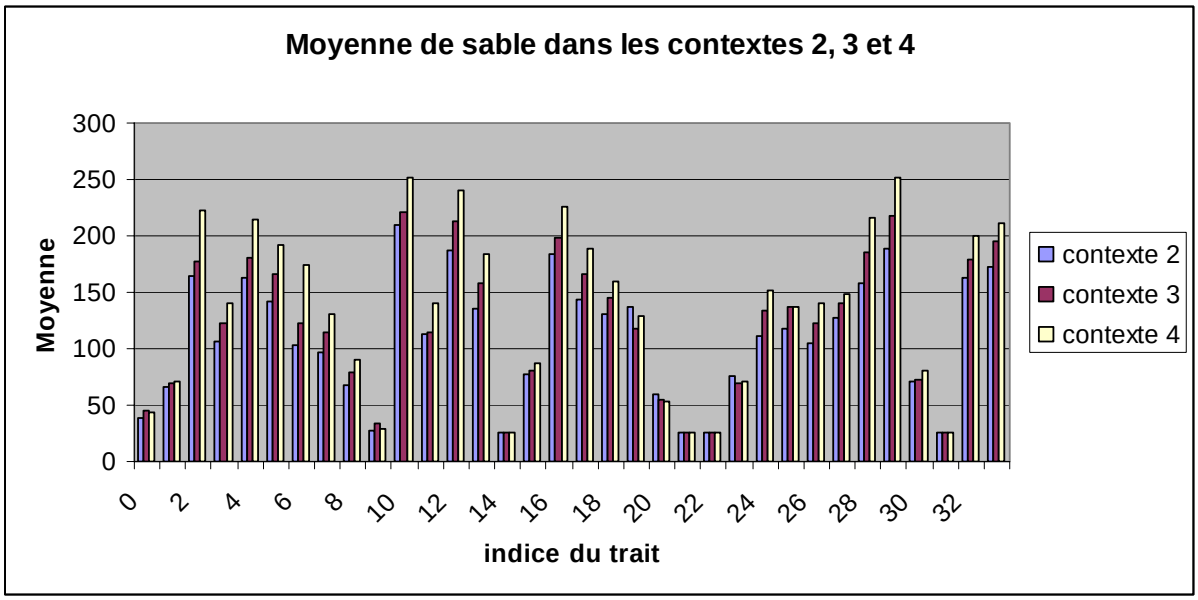
Contexte 3

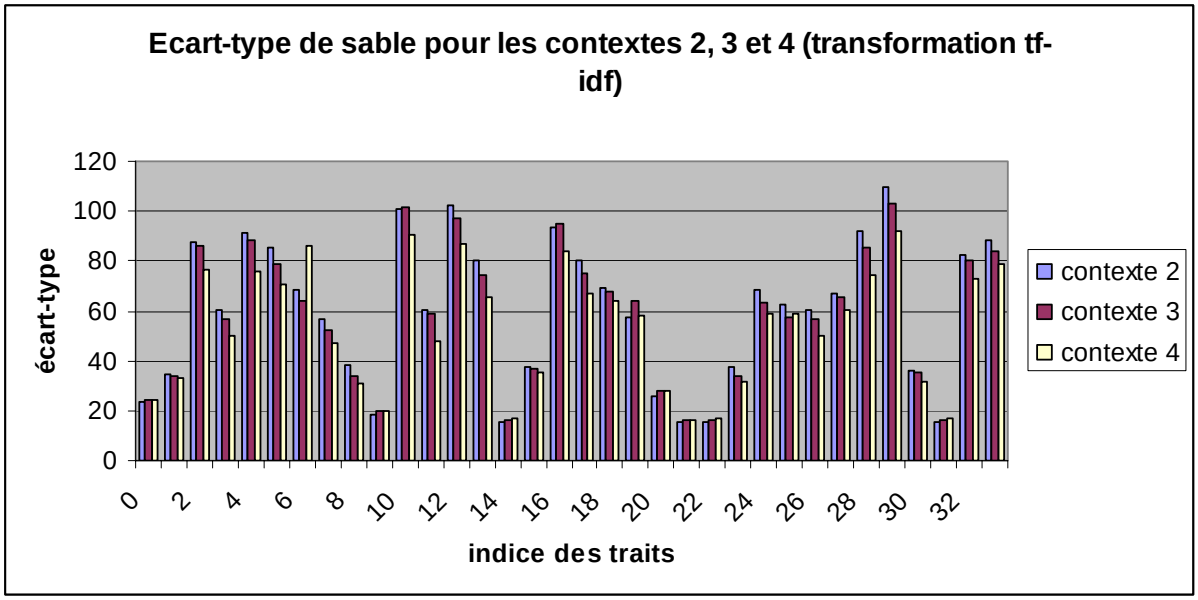


Contexte 4



Les moyennes, quartiles et écarts-types évoluent présentent également la même évolution d'un contexte à l'autre. L'exemple ci-dessous correspond à l'histogramme comparatif des moyennes et écarts-types de sable dans les contextes 2,3 et 4 pour la transformation tf-idf. Des résultats complémentaires sont disponibles en annexe (médiane, quartiles), ainsi que la correspondance entre indices et familles de traits sémantiques (annexes 2 et 5).





Les différentes observations indiquent donc une similarité de comportement des traits sémantiques du mot de référence, aussi bien en moyenne qu'au niveau de leur dispersion. Cette similarité apparaît au niveau des distributions des traits du mot par rapport aux traits sémantiques du contexte et au niveau de la structuration ou des relations entre traits sémantiques du mot.

Nous pouvons avancer plusieurs hypothèses.
D'abord, les phénomènes observés sont peut-être liés aux contextes d'observation choisis : ceux-ci proviennent du corpus de contes et font la taille d'un paragraphe : la taille trop importante du cotexte par rapport à celle du corpus (effet échantillon représentatif) et la nature du cotexte (cotextes issus du corpus lui-même) atténuent probablement les écarts qu'on pourrait mesurer.
Une autre hypothèse est que l'écart-type d'un vecteur-colonne de traits sémantiques est faible par rapport à la moyenne, plus précisément trop faible par rapport aux écarts entre moyennes des différentes colonnes pour faire évoluer la distribution des valeurs de manière significative. L'expérience n°2 a pour but de vérifier cette hypothèse.
Enfin, on peut supposer que la matrice corpus – mot intègre les spécificités du corpus, comme le découpage en sous-unités. La simple sélection d'une sous-matrice, à considérer comme une fonction de projection ou encore une multiplication des coefficients par une indicatrice (valeur 1 si le trait est présent dans le contexte, 0 sinon), ne comporte pas suffisamment d'informations spécifiques au contexte. La fonction appliquée est donc beaucoup trop grossière. Conséquence de ce dernier point : pour des observations à même vocation (observation des sèmes dominants et inhibés dans l'ensemble du sémème ; structuration des sèmes), la matrice contexte – mot peut être considérée comme représentative de la matrice corpus – mot. Autrement dit, l'observation locale est le reflet de l'information globale apportée par le corpus. L'étude des variations fines requiert d'autres outils d'analyse, plus adaptés.

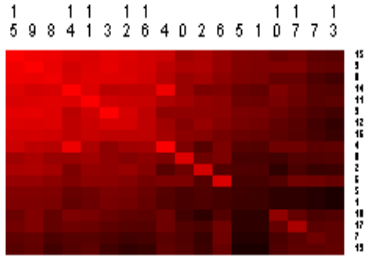
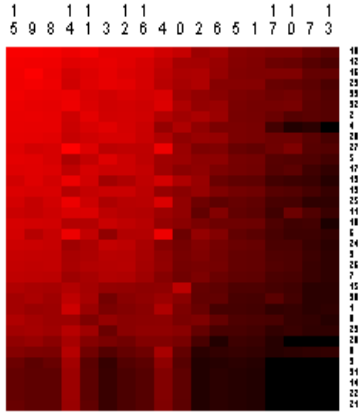
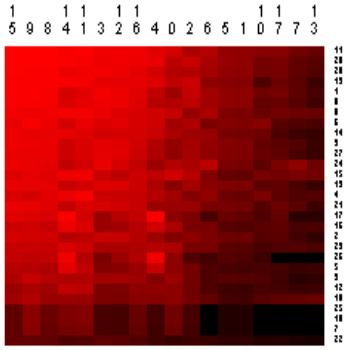
Analyse n°2 : effets de cotextes de taille et de nature différentes

Ces expériences se sont fondées sur des matrices cotexte – mot construites sur le mot pollen et plusieurs types de cotextes :

- des cotextes constitués sur le corpus de Frantext. D'un cotexte à l'autre, les traits activés étaient très différents : activation de /féconder/ chez Proust, /couleur/ chez Schreiber ou encore /grain/ dans certains cotextes de Giraudoux.
- des cotextes constitués d'un seul mot : matrices pollen – pollen, sable – pollen, or –pollen,…

La première série de cotextes devait vérifier l'influence de la nature (genre d'origine, auteur d'origine) du cotexte sur l'agencement des mots. La deuxième série de cotextes a servi à vérifier si la taille des cotextes (en nombre de traits sémantiques) expliquait les résultats précédents.

Voici un exemple de trois matrices mot 1 ~ pollen :

matrice pollen ~ pollen	matrice sable ~ pollen	matrice nacre ~ pollen
		

Liste des traits sémantiques (familles de traits complètes : voir annexe 2) :

0 : fluide	6 : membrane	12 : contenir
1 : être	7 : anthère	13 : utricule
2 : sac	8 : très	14 : grain
3 : loge	9 : petit	15 : fin
4 : poussière	10 : féconder	16 : généralement
5 : produire	11 : jaune	17 : microscopique

On observe que les plages de couleur semblent se répartir de la même manière d'un contexte à l'autre et que les changements d'ordre sont mineurs, avec décalage d'un ou deux rangs. Les constatations faites sur ces exemples particuliers correspondent aux observations sur l'ensemble des tests effectués, aussi bien sur des contextes tirés de Frantext que des contextes réduits à un seul mot.

Dans les deux cas, les résultats ne permettent donc pas de relever de changement majeur concernant l'ordre des traits sémantiques et l'émergence de traits dominants et inhibés. Seules des variations mineures sont observables. L'explication des faibles variations constatées dans la première analyse ne tient donc pas à la nature du contexte, ni à une taille trop importante. C'est pourquoi j'ai exploré la seconde voie d'explication mentionnée au paragraphe précédent : les valeurs des écarts-types par rapport aux écarts entre moyennes.

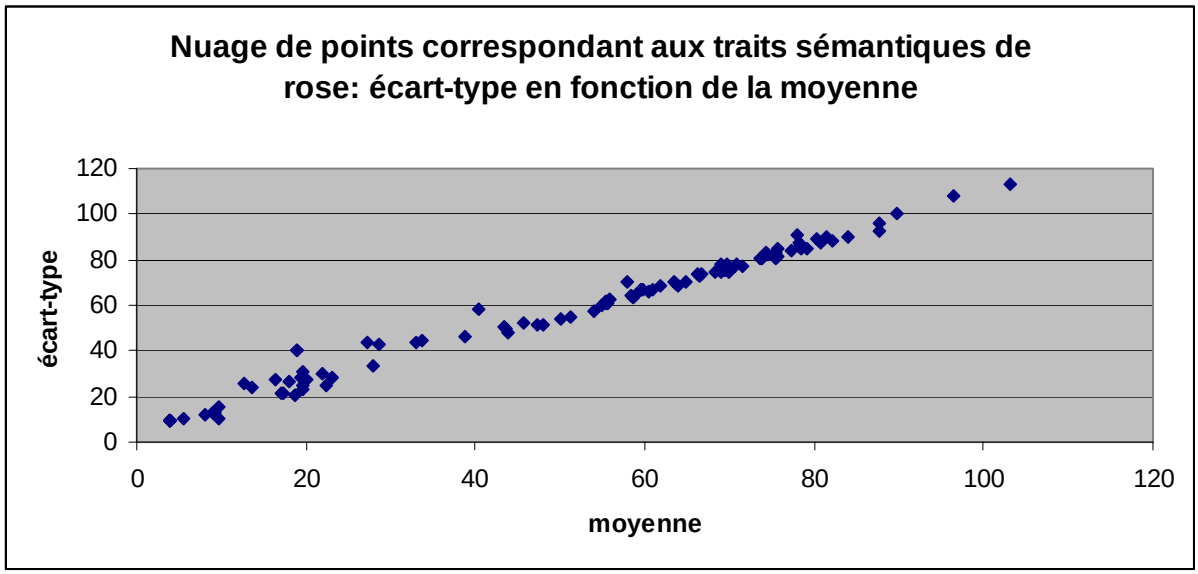
Analyse n°3 : explication de la faible influence des contextes par l'écart-type

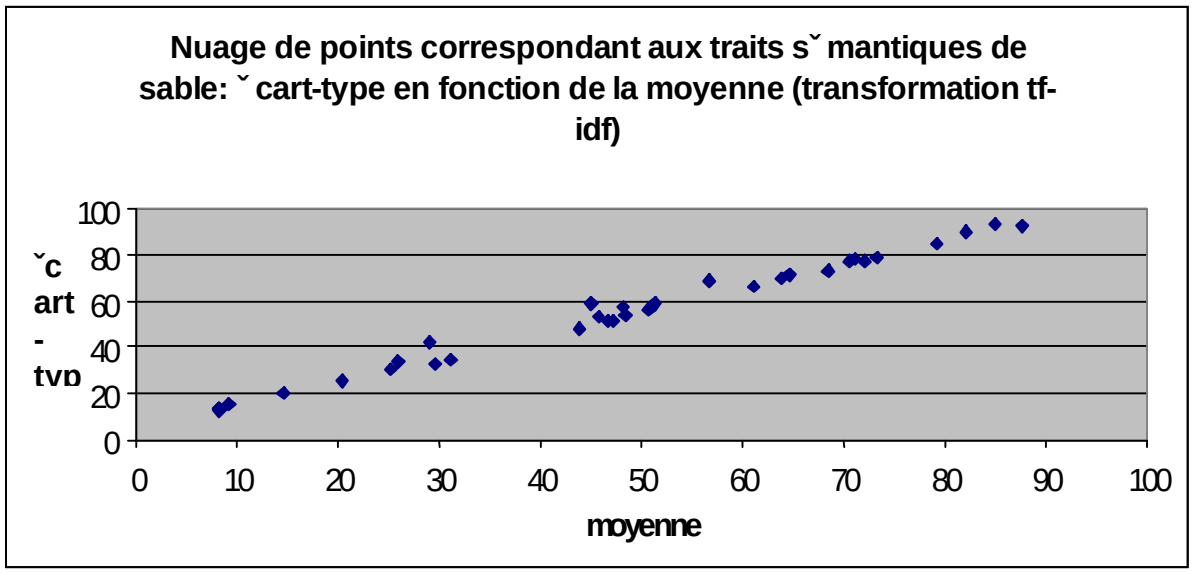
Dans cette analyse, des calculs de moyennes et écarts-types par trait sémantique du mot, puis du rapport écart-type sur moyenne, ont été effectués. Les expériences ont été menées sur le sémème de sable, pollen et rose à partir de la matrice corpus-mot, sur les transformations cosinus, cooccurrences simples (produit scalaire de vecteurs-lignes pour le calcul des cooccurrences), tf-idf et tf-idf suivie de la méthode adaptée de LSA avec 5 valeurs propres puis d'un calcul de cosinus.

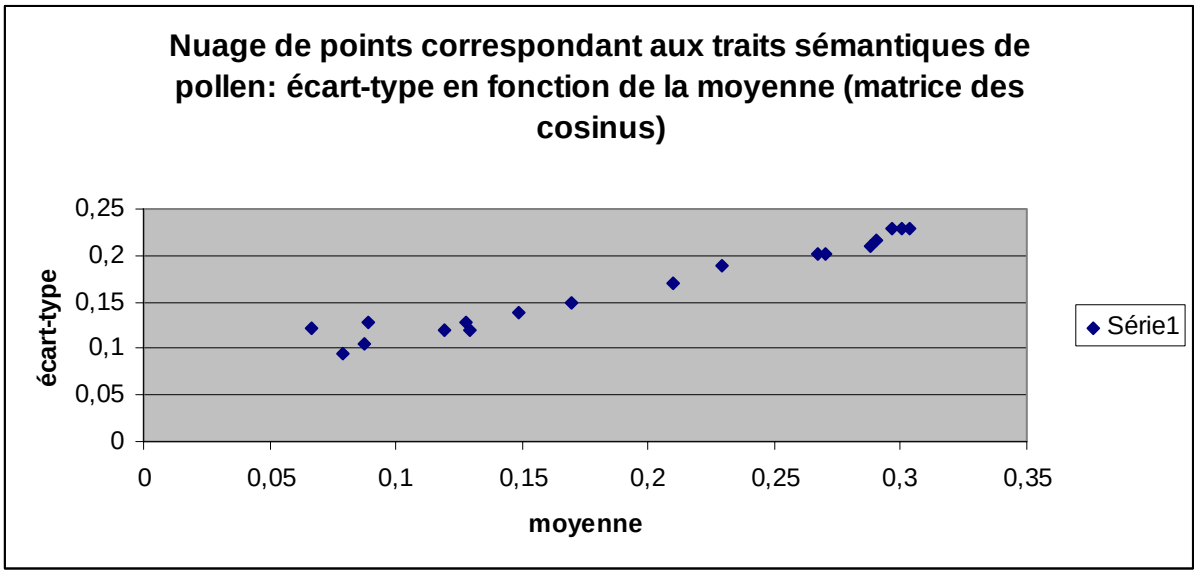
On observe d'une transformation à l'autre des variations assez importantes des rapports écarts-types sur moyenne. Soulignons cependant qu'ils sont assez importants, sauf dans le cas de tf-idf suivie de LSA puis calcul des cosinus. Par exemple, les rapports sont proches de 1 (écart-type de l'ordre de la moyenne) sur la matrice des cosinus de pollen, sable et rose, supérieurs à 1,2 pour la transformation tf-idf.

Une hypothèse explicative de ce phénomène tient à la spécificité des matrices. Dans de nombreux cas, elles peuvent être qualifiées de matrices creuses. Or l'effet des valeurs nulles d'un vecteur sur sa moyenne et son écart-type est non négligeable. Considérons la série de valeurs suivantes : {0 ; 48 ; 52 ; 0 ; 0}. Elle est de moyenne égale à 20 et d'écart-type égal à 24,5. En revanche, la sous-liste composée des coefficients non nuls présente une moyenne de 50 et un écart-type de 2. Non seulement, la présence des zéros baisse fortement la moyenne, mais encore elle contribue à l'augmentation de l'écart-type. Une nouvelle série d'analyse a donc été effectuée sur les vecteurs des traits sémantiques du mot. Leur moyenne et écart-type a été calculé sur une sous-liste des coefficients, à savoir les coefficients non nuls des vecteurs de traits sémantiques. Cependant, les résultats observés n'ont pas été significativement différents des précédents. L'importance de l'écart-type par rapport à la moyenne de chaque trait ne s'explique donc pas seulement par la présence des 0 dans le vecteur.

On observe par ailleurs un autre phénomène assez remarquable, qui, peut-être, ouvre sur une explication du point précédent : dans une représentation avec la moyenne en abscisse et l'écart-type en ordonnée, le nuage de points des traits sémantiques est de forme très allongée, proche d'une droite, comme l'illustrent les exemples ci-dessous, obtenus sur rose, sable et pollen :







Les conclusions, avancées toutefois avec prudence, sont que les traits sémantiques présentent une dispersion similaire. Il se pourrait que ce phénomène soit dû à l'existence de classes sémantiques aux comportements distincts (valeurs dispersées) mais par rapport auxquels deux traits sémantiques évolueront de la même manière. Ces classes sémantiques ne sont pas représentées par deux classes de coefficients distinctes mais par des classes de coefficients échelonnées, ce qui expliquerait la faible différence entre moyennes et écarts-types calculés avec ou sans 0.

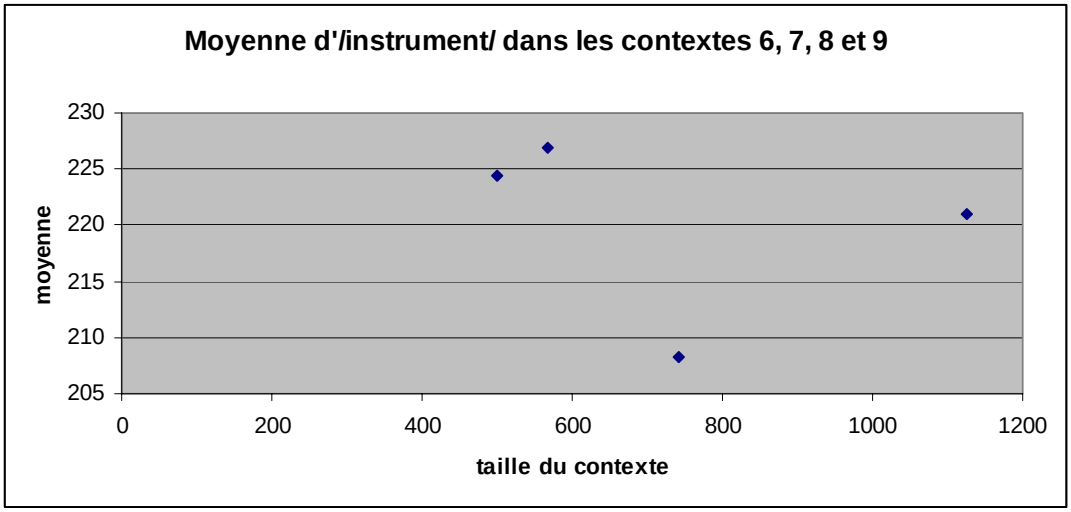
5.3.2.3) Analyse n°4 : mesure des variations fines

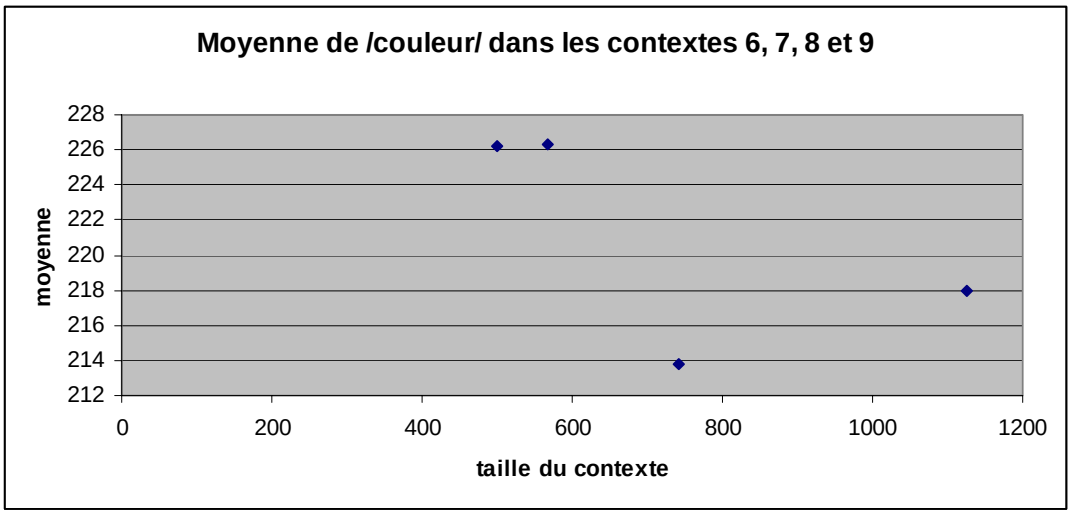
L'objectif de cette analyse est d'étudier l'évolution d'un trait sémantique donné pour une transformation fixée dans différents contextes et de déterminer si un contexte l'active, l'inhibe, ou s'il est dans un état 'neutre'. Pour chaque contexte, le vecteur-colonne correspondant au trait sémantique à étudier est récupéré dans la matrice contexte-mot. D'un contexte à l'autre, les vecteurs n'ont pas la même taille, il est donc nécessaire, pour comparer, de synthétiser les informations contenues dans chaque vecteur par un indicateur commun à tous, la moyenne en l'occurrence.

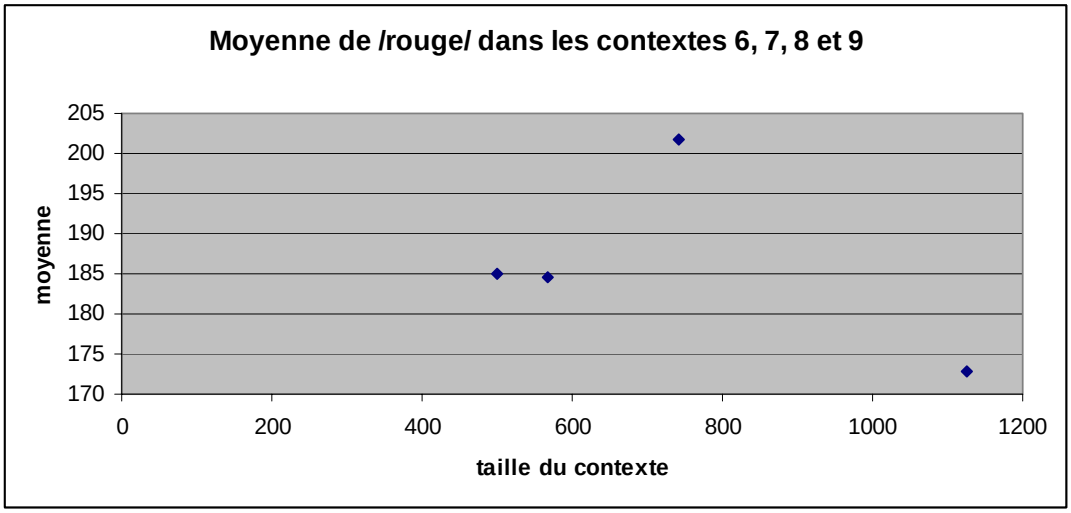
Les expériences ont été effectuées sur des traits sémantiques du mot **rose** : /rouge/, /beauté/, /cœur/, /parfum/, /amour/, /église/, /instrument/ et /vivre/. Les matrices contexte – mot provenant des cooccurrences simples et de tf-idf ont servi de support. Quatre contextes différents ont servi à la comparaison, à savoir les contextes n°6, 7, 8 et 9 disponibles en annexe.

L'indicateur choisi a dans un premier temps été la moyenne des coefficients du vecteur représentatif du trait sémantique. Cependant, une comparaison directe des moyennes implique le présupposé suivant : pour des états d'activation similaires dans différents contextes, les moyennes seront à peu près égales. Or cette hypothèse ne prend pas en compte la taille variable des vecteurs représentatifs des traits sémantiques. Rappelons que cette taille dépend du nombre de traits sémantiques du cotexte. Dans notre cas, les contextes 6, 7, 8 et 9 avaient respectivement des tailles de 741, 1125, 500 et 568 traits sémantiques.

Une comparaison rapide des écarts relatifs aux quatre moyennes n'a révélé aucune relation triviale entre l'agencement relation triviale (par exemple de type linéaire) entre taille du contexte. Les moyennes des cotextes 8 et 9 (plus petites tailles) sont globalement supérieures aux moyennes des cotextes 6 et 7. Les résultats d'/instrument/ et /couleur/ reflètent la tendance généralement observée. Néanmoins, l'allure du nuage de points représentatif des moyennes peut varier de manière importante d'un trait sémantique à l'autre. Il est difficile de savoir quelle est la part d'influence de la spécificité du cotexte sur le plan sémantique et la part des biais dus par exemple à la taille du cotexte.







La solution adoptée consiste à ramener toutes les moyennes à une moyenne de référence, puis à effectuer les comparaisons. Ainsi, pour le cotexte 6 par exemple, la moyenne de tous les traits étudiés (/rouge/, /beauté/, /cœur/, /parfum/, /amour/, /église/, /instrument/ et /vivre/) est divisée par la moyenne d'/église/ dans le cotexte 6. Les moyennes ainsi pondérées semblent donc comparables puisque ramenées à un référentiel commun. Le trait /église/ a été choisi comme référent en raison d'analyses linguistiques préalables, avec identification de l'activation ou non des traits

sémantiques de rose pour chaque contexte. D'après ces analyses, /église/ n'était activé dans aucun des quatre contextes. Ce trait a donc été considéré comme relativement neutre. Cette affirmation doit cependant être considérée avec précaution en raison de la subjectivité de l'interprétation humaine en général, ce qui renvoie à la notion de parcours interprétatif.

Les résultats obtenus sont disponibles en annexe. Le tableau ci-dessous fait le bilan des observations sur la transformation tf-idf :

Trait sémantique	Conformité à l'analyse linguistique	Explication de la non-conformité de certains résultats
amour	-	Le trait apparaît activé dans les contextes 7, 8 et 9 et fortement inhibé dans le contexte 6. Si l'activation pour les contextes 8 et 9 se comprend, l'activation dans le 7 et l'inhibition dans le 8 semblent difficilement pertinents.
beauté	oui	
cœur	oui	
couleur	non	En supprimant le résultat sur le contexte 8, l'agencement des points représentatifs des moyennes est conforme à une analyse linguistique : /couleur/ est fortement activé dans le contexte 7 par rapport au contexte 6, dans lequel le trait est lui-même plus activé que dans le contexte 9. La lecture des différents contextes correspond à ces analyses. En revanche, la forte activation du trait dans le contexte 8 est étonnante et apparaît comme non pertinente.
parfum	oui	
rouge	oui	
vivre	non	/vivre/ recouvre une notion trop générale et, parmi les traits sémantiquement pleins, est un trait à faible contenu sémantique, qui se rapprocherait de la frontière avec le 'sémantiquement vide'.
instrument	non	/instrument/ appartient à une famille de traits sémantiques défectueuse : celle-ci fusionne deux familles, celle d'/installer/ et d'/instrument/. Le programme informatique et l'analyse linguistique (faite uniquement à partir d'/instrument/) ne partent donc pas sur les mêmes bases et peuvent difficilement être conformes, d'autant plus que /installer/ est un terme particulièrement général tandis qu'/instrument/ est assez spécifique.

Les résultats ne sont donc pas tous conformes aux attentes, mais bon nombre de résultats négatifs trouvent une explication cohérente. Les résultats du test peuvent donc être considérés comme relativement probants. Des conclusions plus générales et robustes sur la validité du test demanderaient la réitération des expériences sur un grand nombre de mots et de traits appartenant à chaque mot. Ceci pose cependant le problème de la validation linguistique, réalisée par des humains : si un programme informatique a la capacité de générer en peu de temps une grande masse de données, l'analyse humaine de l'activation ou l'inhibition de traits est une opération longue et requérant de multiples intervenants pour diminuer le phénomène de la subjectivité.

5.3.3) Conclusion sur les expériences

Les différentes expériences ont soulevé un problème majeur : les faibles variations des informations locales par rapport aux informations globales. Le facteur explicatif principal de ces faibles variations n'est ni dû à une faible dispersion, ni à une taille trop importante du cotexte ou à sa

nature. Une explication plausible est que la sélection d'une sous-matrice n'est pas une opération mathématique capable d'appliquer de façon suffisamment importante les spécificités du cotexte local. Les matrices cotexte-mot n'ont donc pas été exploitées pour de la comparaison de cotexte par rapport à l'ensemble du sémème mais elles ont été considérées comme représentatives des informations contenues par la matrice corpus-mot. L'étude des transformations mathématiques sur les matrices cotexte-mot nous informe donc sur la significativité des traits sémantiques d'un sémème telle qu'elle nous apparaît à travers le corpus. La transformation tf-idf est apparue adaptée pour faire émerger des coefficients de significativité pertinents alors que le calcul direct des cooccurrences faisait apparaît comme significatifs des traits à contenu sémantique faible et l'application du χ^2 semblait avoir l'effet inverse, avec une faible significativité pour des termes très fréquents, mais des coefficients de significativité trop importants pour des termes rares. En outre, le passage des occurrences aux cooccurrences a donné de meilleurs résultats avec un produit scalaire entre vecteurs-lignes plutôt qu'un calcul de cosinus. Enfin, pour mesurer l'effet des cotextes, donc l'étude des variations locales à partir de l'image globale, a donné des résultats encourageants avec l'étude d'un trait sémantique donné ramené à une valeur de référence.

Conclusion et perspectives

Le travail effectué a été extrêmement riche et a permis d'explorer des champs très variés. Il a demandé un investissement au niveau théorique aussi bien que pratique. Il a nécessité non seulement de mettre en œuvre des capacités de synthèse d'informations et d'analyse mais aussi de mettre en place une démarche constructive dans la réalisation du programme et la conduite des expériences, ainsi que des efforts pour faire le lien en permanence entre domaine mathématique et domaine linguistique.

Ce stage a néanmoins présenté un certain nombre de difficultés. Il m'a en effet fallu me familiariser avec une discipline inconnue. De plus, la traduction permanente du champ linguistique au champ mathématique, et inversement, s'est avérée être un exercice particulièrement délicat, tant pour la conception du modèle que pour l'interprétation des résultats. De manière plus générale, la communication avec les membres de l'équipe, aux champs de spécialité très divers (différentes approches sémantiques, informatique), s'est révélée indispensable à ma progression et extrêmement enrichissante. Une autre difficulté s'est concrétisée à travers la multiplicité des possibles : les pistes d'exploration étaient extrêmement nombreuses et ont exigé une restriction des ambitions de départ. Entre approfondissement et exploration de nouvelles pistes, j'ai souvent été confrontée à un choix cornélien. Ajoutons comme difficulté les problèmes techniques, c'est-à-dire au niveau informatique : la réalisation d'un programme efficient a été difficile, en particulier à cause des problèmes de manque d'espace mémoire, et n'aurait pas été menée à bien sans l'aide dont j'ai bénéficié.

Pendant quatre mois, le travail réalisé et les réflexions mises en oeuvre m'ont permis de toucher du doigt un terrain extrêmement vaste et presque vierge. Je n'ai pas la prétention d'avoir tout exploré mais pense avoir contribué, très modestement, à quelque avancée dans ce domaine : j'ai ouvert l'une ou l'autre piste, essayé de vérifier la validité d'hypothèses linguistiques au niveau infra-lexical et tenté d'expliciter au niveau linguistique des phénomènes mathématiques ; j'ai proposé un modèle, étudié la validité de celui-ci, sélectionné les voies les plus intéressantes, pistes à creuser ; j'ai enfin généré un outil permettant d'obtenir des données conformément au modèle.

Les perspectives qu'ouvre mon travail se situent à différents niveaux. D'abord, un développement de la plate-forme informatique générée paraît nécessaire : il faudrait résoudre les problèmes d'espace mémoire afin de mener des études sur des corpus de taille plus importante, quitte à recréer des objets Matrice avec une structure plus légère. De plus, la manipulation actuelle du programme est assez artisanale et nécessite une certaine connaissance de celui-ci. Une interface homme-machine (IHM) serait à mettre en place à terme, pour permettre l'exploitation du programme par tout utilisateur.

Sur le plan linguistique, les résultats ont mis à jour des faiblesses dans les regroupements en familles morpho-syntaxiques. Reprendre ceux-ci semble indispensable pour avoir des résultats plus pertinents.

Par ailleurs, le modèle n'en est qu'à des balbutiements, il est loin de refléter finement les phénomènes sémantiques. Plusieurs pistes d'amélioration sont à creuser : il faudrait concevoir un mode de pondération des coefficients afin d'intégrer les spécificités du cotexte, en particulier la syntaxe ; de prendre en compte l'ordre interne aux unités de découpage et entre unités de découpage ; d'intégrer les différentes échelles sémantiques (syntagme, phrase, paragraphe, article pour un corpus journalistique, …). Cette étape requiert le choix et l'étude de fonctions mathématiques appliquées aux paramètres mentionnés. Au niveau de l'analyse, il serait judicieux de réitérer les expériences afin de multiplier les données disponibles et effectuer sur celles-ci des études statistiques plus poussées. L'affinement des outils statistiques à utiliser fait aussi partie des aspects à développer.

Comme dernière perspective, évoquons la conception de nouvelles expériences destinées à repérer des candidats à l'enrichissement du sémème ou encore déterminer des couplages entre traits d'un mot et traits du contexte.

Glossaire

contexte	conditions extralinguistiques d'énonciation et/ou de production d'un texte. Par exemple, l'époque, la pratique sociale (médecine, linguistique, …) correspondant à la production du texte relèvent du contexte.
cooccurrence	apparition conjointe d'unités linguistiques
corpus	ensemble de textes réunis en fonction d'une application particulière
cotexte	voisinage d'un mot pôle dans un texte. Ex : cotexte de 50 mots, paragraphe centré sur un mot.
dédomanialisation	phénomène par lequel un mot se désolidarise de son domaine d'origine. Par ex arpenter était un terme du domaine de la topographie et s'est dédomanialisé pour faire partie de l'usage courant.
désambiguïsation	identification du sens adéquat d'un mot polysémique en fonction du contexte d'apparition
domanialisation	spécialisation d'un mot dans un domaine donné. Ex : clavier dans le domaine informatique
fenêtre de mots	regroupement de mots consécutifs de taille définie
forme sémantique	groupement stable de sèmes structurés
hyperonyme	terme dont le sens inclut celui d'un ou plusieurs autres. Par exemple, rouge est l'hyperonyme de ses hyponymes vermillon, écarlate et cramoisi
hyponyme	antonyme d'hyperonyme
infra-lexical	qui relève des unités linguistiques constitutives du mot. Ex. : les traits sémantiques
lemmatisation	conversion d'une forme en lemme
lemme	forme canonique et conventionnelle d'un mot sous laquelle on range les variations flexionnelles. Ex : le lemme de « mangeront » est « manger » ; le lemme de « petites » est « petit »
lexical	qui relève du mot
molécule sémique	voir forme sémantique
monosémique	qui n'a qu'un seul sens
mot sémantiquement plein	mot dont la valeur relève de son contenu sémantique et non de sa fonction syntaxique ; opposé à mot-outil. Ex : adjectifs, verbes, substantifs, adverbes ; par opposition : déterminants, pronoms
noyau sémique	sous-ensemble du sémème d'un mot instancié de manière récurrente
occurrence	apparition d'une unité linguistique dans un contexte textuel

polysémique	qui a plusieurs sens
regroupement morphologique	regroupement de mots fondé sur leur forme. Ex : grainetier, graine
sémantique	discipline qui étudie le sens
sémantique interprétative	sémantique dont la perspective est herméneutique, c'est-à-dire centrée sur l'interprétation des textes
sémantique textuelle	sémantique des textes, centrée sur l'analyse en composantes textuelles (thématique par exemple) ; assimilée dans ce contexte à la sémantique interprétative
sème	unité minimale de sens ; dans le contexte de mon stage, les sèmes sont extraits automatiquement de définitions lexicographiques
sémème	ensemble des traits sémantiques d'un mot
Sémy	plate-forme d'annotation en traits sémantiques développée par Mick Grzesitchak dans le cadre de ce projet
supra-lexical	qui relève des unités linguistiques supérieures au mot
syntagme	groupe de mots la succession a un sens et qui forment une unité fonctionnelle.
syntaxe	partie de la linguistique qui décrit les règles par lesquelles les unités linguistiques se combinent en syntagmes ou en phrases
TLFi	Trésor de la Langue Française informatisé, dictionnaire de langue développé par l'ATILF
trait sémantique	voir sème

Bibliographie

[Bourion, 2001] Bourion, E., 2001, chapitre 3 de *L'aide à l'interprétation des textes électroniques*, thèse de doctorat, Université Paris X Nanterre

[Caraux & Pinloche, 2005] Caraux, G., Pinloche, S. (2005), « Permutmatrix : A Graphical Environment to Arrange Gene Expression Profiles in Optimal Linear Order, » *Bioinformatics*, **21**, 1280-1281

[Cori & Léon, 2002] Cori, M., Léon, J., 2002, *La constitution du TAL – Etude historique des dénominations et des concepts*, TAL. Volume 43 – n+3-2002, pages 21 à 55. Disponible sur http://infolang.u-paris10.fr/modyco/textes/cori/CoriLeon.PDF.

[Delafosse, 1999] Delafosse, L., 1999, *Glossaire de linguistique computationnelle – Traitement automatique des langues*, http://pagesperso-orange.fr/ldelafosse/Glossaire/Tal.htm

[Dictionnaire des fréquences, 1971] CNRS, Centre de recherche pour un trésor de la langue française, 1971, *Dictionnaire des Fréquences. Études statistiques sur le vocabulaire français. Vocabulaire littéraire des XIX*^{ème} *et XX*^{ème} *siècles. II- Table des fréquences décroissantes*, Nancy

[Habert & Nazarenko, 1997] Habert, B., Nazarenko, A., Salem, A., 1997, « Quantifier les faits langagiers » dans *Les linguistiques de corpus*, Armand Colin/Masson, Paris

[Hatchuel & Tonneau, 1996] Hatchuel, A., Tonneau, S., 1996, chapitre 7 II.2 et III.1, dans *Modèles et décisions statistiques*, Presses de l'Ecole des Mines de Paris.

[Landauer, Foltz & Laham, 1998] Landauer, T.K., Foltz, P.W., Laham, D., 1998, *Introduction to Latent Semantic Analysis (LSA)*. Disponible sur http://lsa.colorado.edu/

[Lemire, 2008] Lemire, « Les lois de Zipf et de Mandelbrot », *Inf6460*, cours en ligne sur la recherche et le filtrage d'informations, disponible sur http://benhur.teluq.uqam.ca/SPIP/inf6460/article.php3?id_article=109&id_rubrique=18

[M. Grzesitchak, 2007] Grzesitchak, M., juin 2007, *Annotation Sémantique de Données Textuelles : Proposition pour l'analyse en traits sémantiques et la recherche d'isotopies*, Master Sciences Cognitives - Spécialité Traitement Automatique des Langues - Université Nancy 2.

[Mauceri, 2007a] Mauceri, C., 2007, « Isotopie et statistiques contextuelles » dans *Indexation et isotopie : vers une analyse interprétative des données textuelles*, Thèse de doctorat

[Mauceri, 2007b] Mauceri, C., Ho, D., 2007, *Clustering By Kernel Density*

[Miller & Torris, 1990] Miller P., Torris T., 1990, *Formalismes syntaxiques pour le traitement automatique du langage naturel*, p. 15, Hermès, Paris

[Missire, 2006] Missire, R., 2006, *Glossaire de sémantique*. Disponible sur http://www.revue-texto.net/Inedits/Missire/Missire_th_glossaire.pdf

[Muller, 1968] Muller, C., 1968, *Initiation à la statistique linguistique*, Larousse

[Pierrel, 1997] Pierrel, J.-M., 1997, *Ingénierie des langues*, Hermès

[Ramdani, 2007] Ramdani, E., 2007, *Du dictionnaire de langue au lexique TAL – la construction d'une ressource pour l'annotation sémantique des textes*, Mémoire de Master

[Rastier, 1996] Rastier, F., 1996, *La sémantique des textes : concepts et applications*, Texto !. Disponible sur http://www.revue-texto.net/Inedits/Rastier/Rastier_Concepts.html

[Rouchaleau, 2008] Rouchaleau, Y., 2008 , p.18, *Traitement numérique du signal*, Les Presses de l'Ecole des Mines

[Valette, 2004] Valette, M., 2004, *Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur Internet, Approches Sémantiques du Document Numérique*, Actes du 7e Colloque International sur le Document Electronique, 22-25 juin 2004, Patrice Enjalbert et Mauro Gaio, eds, pp.215-230

[Valette, 2006c] Valette, M., 2001, « Observations sur la nature et la fonction des emprunts conceptuels en sciences du langage », Corpus en Lettres et Sciences sociales : *des documents numériques à l'interprétation*, Actes du colloque international d'Albi, juillet 2006. Carine Duteil, Baptiste Foulquié (publ.), François Rastier, Michel Ballabriga (dir.), Paris, Texto, 2006. ISSN 1773-0120

[Valette, Estacio-Moreno, Petitjean & Jacquey, 2006] Valette, M., Estacio-Moreno, A., Petitjean, E., Jacquey, E., *Eléments pour la génération de classes sémantiques à partir de définitions lexicographiques pour une approche sémique du sens*, paru dans *Verbum ex Machina,Actes de la 13e conférence sur le traitement automatique des langues naturelles (TALN06)*,Piet Mertens, Cédrick Fairon, Anne Dister, Patrick Watrin (éds), *Cahier du CENTAL*, 2.1, UCL Presses Universitaires de Louvain. Volume 1. Pages 357-366.

[Valette & Grabar, 2004] Valette, M., Grabar, N., 2004, *Caractérisation de textes à contenu idéologique : statistique textuelle ou extraction de syntagme ? L'exemple du projet PRINCIP, Le poids des mots*, Actes des 7e Journées internationales d'Analyse statistique des Données Textuelles (JADT), Louvain-la-Neuve (Belgique)

[Valette & Rastier, 2008] Valette, M., Rastier, F., à paraître (sept 2008), *De la polysémie à la néosémie*

[Venant, 2004] Venant, F., 2004, *Géométriser le sens*, Actes de la conférence RECITAL , Fès, Maroc

[Victorri, 1994] Victorri, B., 1994, *The use of continuity in modelling semantic phenomena*

[Victorri, 2005] Victorri, B., 2005, « Polysémie lexicale », dans *Sémantique et traitement automatique du langage naturel*, chapitre, Hermès

ANNEXES

A1) Code informatique, éléments principaux du programme réalisé en Java

Classe principale (sans le main) : ReprSem0

```
import java.text.NumberFormat;
import java.io.*;
import java.util.Scanner;
import java.util.regex.*;
import java.util.*;
import java.lang.Integer;
import java.sql.SQLException;

public class ReprSem0 {

  public static void main(String [] args) {
  }

  // Méthode pour récupérer sous forme de table de hachage les données
fournies par un fichier de SEMY
  // Entrées : nombre d'unités de découpage, fichier source
  // sortie : table de hachage contenant
  // 1) comme clé le numéro de la famille de sèmes,
  // 2) comme valeur l'objet SemEtDistri composé de {famille de sèmes
(attribut 1)+ distribution (attribut 2)}
  public static HashMap<Integer,SemEtDistri> lireFichierSemy(int
nombreDeContextes, File fichierSource) {
    int nbCtx = nombreDeContextes;
    // Les expressions régulières à  repérer dans le fihier de produit par
Sémy :
    // 1) item : les sèmes
    // 2) distri : la distribution sous forme de {0: 1, 1:3},...
    // 3) les indicateurs de famille de sèmes : /1204/ par ex (correspond
à  la famille 1204 d'Egle)
    Pattern item=Pattern.compile("\\-?\\p{L}+\\-?\\p{L}*(,\\-?\\p{L}+\\-
?\\p{L}*)*", Pattern.MULTILINE);
    Pattern distri = Pattern.compile("\\{(\\d+: \\d+, )*(\\d+: \\d+)\\}",
Pattern.MULTILINE);
    Pattern familleSM = Pattern.compile("\\/\\d+\\/");
    HashMap<Integer,SemEtDistri> tbSem = new HashMap<Integer,SemEtDistri>();
    // Utilisation d'un Scanner qui parcourt le texte
    try {
      Scanner sc = new Scanner(fichierSource,"UTF-8");
      for (int i=0 ; i<6 ; i++) {
        sc.nextLine();
      }

      do {
        ArrayList<String> itemSemes = new ArrayList<String>(); // Liste
contenant tous les sèmes de la famille de sèmes
        double [] distriOcc = new double[nbCtx];
        String number = sc.findWithinHorizon("\\d+",Pattern.UNICODE_CASE);

        // Si le scanner trouve un nombre dans la ligne (numéro indiqué en
début de ligne):
```

```
        if (number != null) {
          int num = Integer.parseInt(number);
          System.out.println("Lecture ligne "+num);

          // Sauter les indicateurs des familles d'Egle
          if (sc.hasNext(familleSM)) {
            itemSemes.add(sc.next());
            sc.next();
          }

          // génération de la liste de sèmes composant la famille de sèmes
          while (sc.hasNext(item)) {
            itemSemes.add(sc.findInLine(item));
          }
          if (itemSemes.isEmpty()) {
            itemSemes.add("Mot non identifie, voir item numéro "+num);
          }

          // Récupération de la répartition sous forme d'un tableau de
double (nième cas : fenêtre n ; valeur : nb d'occurrences dans la fenêtre n)
          String val=sc.findInLine("\\d+");

          int numCol=0;
          double nbOcc=0;
          boolean col=true;

          while (val!=null)  {
            if (col) {
              numCol = Integer.parseInt(val);
              col = false;
            }
            else {
              nbOcc = Double.parseDouble(val);
              distriOcc[numCol] = nbOcc;
              col = true;
            }
            val=sc.findInLine("\\d+");
          }
          SemEtDistri sd = new SemEtDistri(itemSemes,distriOcc);
          tbSem.put(num,sd);
          sc.next("\\}"); // Sans cette ligne, boucle while infinie
à  cause de l'accolade
        }
        else return tbSem;
      } while (sc.hasNextLine());
      sc.close();
    }
    catch (Exception e){
      System.out.println(e.toString());
    }
    return tbSem;
  }


  //Méthode pour récupérer une matrice à  partir de la table de hachage
  // Entrée : la table de hà¢chage dont on veut récupérer les coefficients ;
le nombre de contextes en lesquels le corpus a été découpé
  // Sortie : la matrice d'occurrences correspondant à  la table ; nombre
de lignes = nombres de familles de sèmes ; nombre de colonnes = nombre de
contextes
  public static Matrice toMatrice(HashMap<Integer,SemEtDistri> tbSem) {
```

```
    int nbLgn=tbSem.size();
    System.out.println("nb de lignes de toMatrice : "+nbLgn);
    int nbCol = ReprSem0.getNbCtx(tbSem);
    System.out.println("nb de colonnes de toMatrice : "+nbCol);
    Matrice mat = new Matrice(nbLgn,nbCol);
    for (int i=0 ; i<nbLgn ; i++) {
      double [] valeurs=tbSem.get(i).getOcc();
      int nC=valeurs.length;
      for (int j=0 ; j<nC ; j++) {
        mat.set(i,j,valeurs[j]);
      }
    }
    return mat;
  }

  // Méthode pour reconstituer une table de hachage après transformations
mathématiques
  // Entrée : la matrice ayant subi les transformation ; la table de
hachage d'origine dont les valeurs (distributions) ont été transformées
  /// Sortie : la table de hachage avec les nouvelles valeurs de
distribution
  public static HashMap<Integer,SemEtDistri>
remplaceDistri(HashMap<Integer,SemEtDistri> tbSem, Matrice mat) {
    HashMap<Integer,SemEtDistri> newTb = new HashMap<Integer,SemEtDistri>();
    double [][] tbMat = mat.getArray();
    double [] lgnMat;
    SemEtDistri sed;
    SemEtDistri sed2;
    for (int i=0 ; i<tbMat.length ; i++) {
      lgnMat = tbMat[i];
      sed=tbSem.get(i);
      sed2 = new SemEtDistri(sed.getSemes(),lgnMat);
      newTb.put(i,sed2);
    }
    return newTb;
  }

  // Méthode pour sélectionner seulement certaines entrées de la tabe de
hachage
  // les entrées sélectionnées sont les sèmes contenus dans une autre table
de hachage
  public static HashMap<Integer,SemEtDistri>
getLines(HashMap<Integer,SemEtDistri> tbSupport,
HashMap<Integer,SemEtDistri> tbAExtraire) {
    HashMap<Integer,SemEtDistri> tbExtraite = new
HashMap<Integer,SemEtDistri>();
    int sizeSup = tbSupport.size(); // nombre de familles de sèmes du
corpus
    System.out.println(" Taille de tbSupport : "+sizeSup);
    Set<Integer> indicesTax = tbAExtraire.keySet();// liste des numéro des
sèmes à  identifier
    int nbCles = indicesTax.size();
    System.out.println("Nombre de sèmes à  extraire : "+nbCles);

    boolean [] cleIdentifiee = new boolean[nbCles];
    int nbIdentifie = 0;

    for (int i=0 ; i<sizeSup ; i++) { //en parcourant la table de hachage
du corpus
      SemEtDistri sed = tbSupport.get(i); //récupère le ième sème du corpus
et sa distribution
```

```
      ArrayList<String> smSup = sed.getSemes(); // n'en garde que le sème

      for (int j : indicesTax) { //en parcourant les entrées des sèmes du
mot / contexte pas encore reliés aux sèmes du corpus
        if (cleIdentifiee[j]==true) {
          continue;
        }
        else {
          ArrayList<String> smInf = tbAExtraire.get(j).getSemes();
//récupère le sème
          if (smSup.get(0).equals(smInf.get(0))) { // si le sème du corpus
et du mot/contexte sont identiques
            tbExtraite.put(j,sed); // prend dans la table du corpus le sème
et sa distri, mets-les dans  la nouvelle table de hachage
            // System.out.print("lgn = "+j+", val="+sed.getOcc()[0]+",
nbCol="+sed.getOcc().length+" ; ");
            cleIdentifiee[j]=true;
            nbIdentifie++;
            if (nbIdentifie==nbCles) {
              System.out.println("Nbidentifie == nbCles");
//          if (indicesTax.isEmpty()) { // si la liste est vide
              return tbExtraite; // retourne la table de hachage
            }
            break;
          }
        }
      }
    }
    int nbColonnes = tbSupport.get(0).getOcc().length;
    for (int j:indicesTax) {
      if (cleIdentifiee[j]==false) {
        SemEtDistri sed = tbAExtraire.get(j);
        tbExtraite.put(j,sed);
      }
    }
    return tbExtraite;
  }

  public static int getNbCtx(HashMap<Integer,SemEtDistri> hm) {
    int nbCtx=0;
    if (!hm.isEmpty()) {
      nbCtx=hm.get(0).getOcc().length;
    }
    return nbCtx;
  }
  // Méthode pour obtenir la table de hachage de la transposée de la
matrice sèmes du corpus - sèmes du mot
  // ATTENTION : la matrice du corpus doit être symétrique
  // correspond à  une sélection de colonnes
  // Entrée :
  // 1) la table de hachage dont les distributions vont être transformées
(sous-tableau du tableau de distri)
  // 2) la table de hachage permettant de sélectionner le bon sous-tableau
  // 3) le nombre de contextes, ie la taille des tableaux de distribution
dans la table de hà¢chage dont on ne va garder qu'une partie
  // Sortie : La nouvelle table de hà¢chage avec pour distribution le sous-
tableau du tableau initial (sélectionné en fonction des sèmes de la table
de hà¢chage servant à  la sélection)
  public static HashMap<Integer,SemEtDistri>
getColumns(HashMap<Integer,SemEtDistri> tbCorpus,
HashMap<Integer,SemEtDistri> tbMot) {
```

```
    HashMap<Integer,SemEtDistri> tbInterm
=ReprSem0.getLines(tbCorpus,tbMot);
    Matrice mTransposee = ReprSem0.toMatrice(tbInterm);
    Matrice me = mTransposee.transpose();
    return ReprSem0.remplaceDistri(tbCorpus,me);
  }

// Méthode pour sérialiser les objets matrices
  public static void serialiser(Object o, String nomFichier) {
    try {
      System.out.println("Serialisation en cours");
      FileOutputStream fos = new FileOutputStream(nomFichier);
      ObjectOutputStream oos = new ObjectOutputStream(fos);
      try {
        oos.writeObject(o);
        oos.flush();
        Matrice matVide=new Matrice();
        boolean matrice=o.getClass().isInstance(matVide);
        HashMap<Integer,SemEtDistri> hmVide = new
HashMap<Integer,SemEtDistri>();
        boolean hashMap = o.getClass().isInstance(hmVide);
        if (matrice) {
          System.out.println("Matrice serialisee");
        }
        else if (hashMap) {
          System.out.println("Table de hachage serialisee");
        }
      } finally {
        try {
          oos.close();
        } finally {
          fos.close();
        }
      }

    } catch (IOException ioe) {
      ioe.printStackTrace();
    }

  }

  // Méthode de désérialisation d'une Matrice
  public static void deserialiserMatrice (String nomFichier){
    try {
      // ouverture d'un flux d'entrée depuis le fichier nomFichier
      FileInputStream fis = new FileInputStream(nomFichier);
      // création d'un "flux objet" avec le flux fichier
      ObjectInputStream ois= new ObjectInputStream(fis);

      try {
        // désérialisation : lecture de l'objet depuis le flux d'entrée
        Matrice mat=(Matrice)ois.readObject();
        if (mat.getRowDimension() != 0 && mat.getColumnDimension() !=0) {
          System.out.println("Matrice deserialisee");
        }


      } finally {
        // on ferme les flux
        try {
          ois.close();
```

```
        } finally {
          fis.close();
        }
      }
    } catch(IOException ioe) {
      ioe.printStackTrace();
    } catch(ClassNotFoundException cnfe) {
      cnfe.printStackTrace();
    }
  }

  // Méthode de désérialisation d'une HashMap
  public static void deserialiserHashMap (String nomFichier){
    try {
      // ouverture d'un flux d'entrée depuis le fichier nomFichier
      FileInputStream fis = new FileInputStream(nomFichier);
      // création d'un "flux objet" avec le flux fichier
      ObjectInputStream ois= new ObjectInputStream(fis);

      try {
        // désérialisation : lecture de l'objet depuis le flux d'entrée
        HashMap<Integer,SemEtDistri>
hMap=(HashMap<Integer,SemEtDistri>)ois.readObject();
        if (hMap!=null) {
          System.out.println("Table de hachage deserialisee");
        }

      } finally {
        // on ferme les flux
        try {
          ois.close();
        } finally {
          fis.close();
        }
      }
    } catch(IOException ioe) {
      ioe.printStackTrace();
    } catch(ClassNotFoundException cnfe) {
      cnfe.printStackTrace();
    }
  }

  // Pour récupérer des fichiers textes avec les valeurs des matrices
  public static void exporterFichier(Matrice mat, String nomFichiercsv) {
    try {
      PrintWriter out = new PrintWriter(nomFichiercsv);
      StringBuilder buffer = new StringBuilder(32*1024*1024);
      int nbLgn = mat.getRowDimension();
      int nbCol = mat.getColumnDimension();

      buffer.append("indices\t");
      for (int i=0 ; i<nbCol ; i++) {

        buffer.append(i);
        buffer.append("\t");
      }
      buffer.append("\n");


      int n=0;
      int c=0;
```

```
      for (int i=0 ; i<nbLgn ; i++) {
        n++;
        buffer.append(i);
        buffer.append("\t");
        int n2=0;
        for (int j=0 ; j<nbCol ; j++) {
          double val = Math.floor(mat.get(i,j)*100)/100;
          double nbTronque = val;
          n2++;
          buffer.append(nbTronque);
          buffer.append("\t");
        }
        buffer.append("\n");
      }

      out.append(buffer);
      buffer.replace(0,buffer.length()-1,"");
      out.close();
    }
    catch (IOException e) {
      System.out.println(e);
    }
  }

  public static void exporterFichier(HashMap<Integer,SemEtDistri> al,
String nomFichier) {
    Matrice mat = ReprSem0.toMatrice(al);
    ReprSem0.exporterFichier(mat, nomFichier);
  }
}
```

Classe SemEtDistri

```
import java.util.*;
import java.io.Serializable;

public class SemEtDistri implements Serializable {
  private static final long serialVersionUID = 70L;
  ArrayList<String> grSemes; // sème et famille (d'après les rgpts d'Egle)
à  laquelle il appartient
  double [] nbOccU; // nombre d'occurrences unitaire, ie par dÃ©coupage ;
entrées d'une ligne de la matrice d'occurrences ou cooccurrences

  public SemEtDistri(ArrayList<String> al, double [] d) {
    grSemes = new ArrayList<String>(al);
    nbOccU = new double [d.length];
    System.arraycopy(d,0,nbOccU,0,d.length);
  }

  public ArrayList<String> getSemes() {
    return grSemes;
  }
  public String getSemes(int i) {
    int taille = grSemes.size();
    if (taille == 0) {
      String s = "Mot avec accent, pb de lecture";
      return s;
    }
    else if (i<taille) {
```

```
      return grSemes.get(i);
    }
    else {
      System.out.println("Il n'y a que "+taille+" sèmes dans la groupement.
Récupération du premier sème.");
      return grSemes.get(0);
    }
  }

  public double [] getOcc() {
    return nbOccU;
  }
}
```

Classe Matrice

Remarque : seules sont présentes les méthodes ajoutées à la classe Matrix du package Jama disponible sur http://math.nist.gov/javanumerics/jama/

```
   /** Somme les coefficients d'une colonne
   @param c    indice de colonne
   @return     sum, somme des éléments de la colonne
   @exception  ArrayIndexOutOfBoundsException
   */

   public double sumCol(int c)
   {
     double sum=0;
     for (int j=0;j<m;j++){
       sum+=A[j][c];
     }
     return sum;
   }

   /** Somme les coefficients d'une ligne
   @param c    indice de ligne
   @return     sum, somme des éléments de la ligne
   @exception  ArrayIndexOutOfBoundsException
   */

    public double sumLgn(int l){
      double sum=0;
      for (int j=0 ; j<n;j++){
        sum+=A[l][j];
      }
      return sum;
    }

    /** Somme tous les coefficients de la matrice
   @return     sum, somme de tous les coefficients
   */
    public double sumTot() {
      double sum=0;
      for (int i=0 ; i<m;i++){
        for (int j=0 ; j<n ; j++) {
          sum+=A[i][j];
        }
      }
      return sum;
    }
```

```
/** Moyenne par ligne pour toutes les lignes de la matrice
  @return d, vecteur (1,m) des moyennes
  */
public double [] moyLgns() {
  double [] d=new double[m];
  for (int i=0 ; i<m ; i++) {
    d[i]=0;
    for (int j=0 ; j<n ; j++) {
      d[i]+=A[i][j];
    }
    d[i]=d[i]/n;
  }
  return d;
}

/**Moyenne par colonne pour toutes les colonnes de la matrice
  @return d, vecteur (1,n) des moyennes
  */
public double [] moyCols() {
  double [] d=new double[n];
  for (int j=0 ; j<n ; j++) {
    d[j]=0;
    for (int i=0 ; i<m ; i++) {
      d[j]+=A[i][j];
    }
    d[j]=d[j]/m;
  }
  return d;
}

/**Calcul de l'écart-type des lignes
  @param moy : vecteur des moyennes par ligne
  @return d : vecteur des écarts-types par ligne
  */
public double [] sigmaLgns(double [] moy) {
  double [] d=new double[m];
  for (int i=0 ; i<m ; i++) {
    d[i]=0;
    for (int j=0 ; j<n ; j++) {
      d[i]+=A[i][j]*A[i][j];
    }
    d[i]=Math.sqrt(d[i]/n-moy[i]*moy[i]);
  }
  return d;
}

/**Calcul de l'écart-type des colonnes
  @param moy : vecteur des moyennes par colonne
  @return d : vecteur des écarts-types par colonne
  */
public double [] sigmaCols(double [] moy) {
  double [] d=new double[n];
  for (int j=0 ; j<n ; j++) {
    d[j]=0;
    for (int i=0 ; i<m ; i++) {
      d[j]+=A[i][j]*A[i][j];
    }
    d[j]=Math.sqrt(d[j]/m-moy[j]*moy[j]);
  }
  return d;
```

```
    }

    /**Calcul de l'écart-type des lignes
      @return d : vecteur des écarts-types par ligne
      */
    public double [] sigmaLgns() {
      double [] s=new double[m];
      double [] moy=new double[m];
      for (int i=0 ; i<m ; i++) {
        s[i]=0;
        moy[i]=0;
        for (int j=0 ; j<n ; j++) {
          moy[i]+=A[i][j];
          s[i]+=A[i][j]*A[i][j];
        }
        moy[i]=moy[i]/n;
        s[i]=Math.sqrt(s[i]/n-moy[i]*moy[i]);
      }
      return s;
    }

    /**Calcul de l'écart-type des colonnes
      @return d : vecteur des écarts-types par colonne
      */
    public double [] sigmaCols() {
      double [] s=new double[n];
      double [] moy=new double[n];
      for (int j=0 ; j<n ; j++) {
        s[j]=0;
        moy[j]=0;
        for (int i=0 ; i<m ; i++) {
          moy[j]+=A[i][j];
          s[j]+=A[i][j]*A[i][j];
        }
        moy[j]=moy[j]/m;
        s[j]=Math.sqrt(s[j]/m-moy[j]*moy[j]);
      }
      return s;
    }

    /**Moyenne par ligne pour toutes les lignes de la matrice sans compter
les 0
      @return d, vecteur (1,n) des moyennes des coefs non nuls par ligne
      */
    public double [] moyLgnsSans0() {
      double [] d=new double[m];
      int nbCoefNonNuls;
      for (int i=0 ; i<m ; i++) {
        d[i]=0;
        nbCoefNonNuls=0;
        for (int j=0 ; j<n ; j++) {
          if (A[i][j]!=0) {
            d[i]+=A[i][j];
            nbCoefNonNuls+=1;
          }

        }
        if (nbCoefNonNuls!=0) {
          d[i]=d[i]/nbCoefNonNuls;
        }
      }
```

```
      return d;
    }
    /**Moyenne par colonne pour toutes les colonnes de la matrice sans
compter les 0
      @return d, vecteur (1,n) des moyennes des coefs non nuls par colonne
      */
    public double [] moyColsSans0() {
      double [] d=new double[n];
      int nbCoefNonNuls;
      for (int j=0 ; j<n ; j++) {
        nbCoefNonNuls=0;
        d[j]=0;
        for (int i=0 ; i<m ; i++) {
          if (A[i][j]!=0) {
            d[j]+=A[i][j];
            nbCoefNonNuls+=1;
          }
        }
        if (nbCoefNonNuls!=0) {
          d[j]=d[j]/nbCoefNonNuls;
        }
      }
      return d;
    }

    /**Calcul de l'écart-type des lignes sans les 0
      @param moy : vecteur des moyennes par ligne sur les coefs non nuls
      @return d : vecteur des écarts-types par ligne sur les coefs non nuls
      */
    public double [] sigmaLgnsSans0(double [] moy) {
      int nbCoefNonNuls;
      double [] d=new double[m];
      for (int i=0 ; i<m ; i++) {
        nbCoefNonNuls=0;
        d[i]=0;
        for (int j=0 ; j<n ; j++) {
          if (A[i][j]!=0) {
            d[i]+=A[i][j]*A[i][j];
             nbCoefNonNuls+=1;
          }
        }
        if (nbCoefNonNuls!=0) {
          d[i]=Math.sqrt(d[i]/nbCoefNonNuls-moy[i]*moy[i]);
        }
      }
      return d;
    }

    /**Calcul de l'écart-type des colonnes sans les 0
      @param moy : vecteur des moyennes par colonne sur les coefs non nuls
      @return d : vecteur des écarts-types par colonne sur les coefs non
nuls
      */
    public double [] sigmaColsSans0(double [] moy) {
      int nbCoefNonNuls;
      double [] d=new double[n];
      for (int j=0 ; j<n ; j++) {
        nbCoefNonNuls=0;
        d[j]=0;
        for (int i=0 ; i<m ; i++) {
          if (A[i][j]!=0) {
```

```
            d[j]+=A[i][j]*A[i][j];
             nbCoefNonNuls+=1;
          }
        }
        if (nbCoefNonNuls!=0) {
          d[j]=Math.sqrt(d[j]/nbCoefNonNuls-moy[j]*moy[j]);
        }
      }
      return d;
    }

    /**Calcul de l'écart-type des lignes sans les 0
      @return d : vecteur des écarts-types par ligne sur les coefs non nuls
      */
    public double [] sigmaLgnsSans0() {
      int nbCoefNonNuls;
      double [] s=new double[m];
      double [] moy=new double[m];
      for (int i=0 ; i<m ; i++) {
        nbCoefNonNuls=0;
        s[i]=0;
        moy[i]=0;
        for (int j=0 ; j<n ; j++) {
          if (A[i][j]!=0) {
            moy[i]+=A[i][j];
            s[i]+=A[i][j]*A[i][j];
            nbCoefNonNuls+=1;
          }
        }
        if (nbCoefNonNuls!=0) {
          moy[i]=moy[i]/nbCoefNonNuls;
          s[i]=Math.sqrt(s[i]/nbCoefNonNuls-moy[i]*moy[i]);
        }
      }
      return s;
    }

    /**Calcul de l'écart-type des colonnes sans les 0
      @return d : vecteur des écarts-types par colonne sur les coefs non
nuls
      */
    public double [] sigmaColsSans0() {
      int nbCoefNonNuls;
      double [] s=new double[n];
      double [] moy=new double[n];
      for (int j=0 ; j<n ; j++) {
        nbCoefNonNuls=0;
        s[j]=0;
        moy[j]=0;
        for (int i=0 ; i<m ; i++) {
          if (A[i][j]!=0) {
            moy[j]+=A[i][j];
            s[j]+=A[i][j]*A[i][j];
            nbCoefNonNuls+=1;
          }
        }
        if (nbCoefNonNuls!=0) {
          moy[j]=moy[j]/nbCoefNonNuls;
          s[j]=Math.sqrt(s[j]/nbCoefNonNuls-moy[j]*moy[j]);
        }
      }
```

```
      return s;
    }

   /** Multiplication matricielle d'une matrice et de sa transposee, A * B'
   @param B    another matrix
   @return     Matrice product, A * B'
   @exception  IllegalArgumentException Matrice inner dimensions must agree.
   */

   public Matrice timesTranspose (Matrice B) {
      if (B.n != n) {
         throw new IllegalArgumentException("Matrice inner dimensions must
agree.");
      }
      Matrice X = new Matrice(m,B.m);
      double[][] C = X.getArray();
      for (int j = 0; j < B.m; j++) {
         for (int i = 0; i < m; i++) {
            double s = 0;
            for (int k = 0; k < n; k++) {
               s += A[i][k]*B.A[j][k];
            }
            C[i][j] = s;
         }
      }
      return X;
   }

/* ------------------------
   Mes Methodes
 * ------------------------ */

   /** Calcul du nombre d'occurrences moyen, fondé sur le principe de la
distance du chi2 (hyppothèse d'indépendance)
    @return M, matrice des coefficients moyens
    */

   public Matrice chi2() {
     Matrice M=new Matrice(m,n);
     double [][] X=M.getArray();
     double ni;
     double nj;
     double nT=sumTot();
     for (int i=0 ; i<m ; i++) {
       ni=sumLgn(i);
       for (int j=0 ; j<n ; j++) {
         nj=sumCol(j);
         if (nT!=0) {
           X[i][j]=ni*nj/nT;
         }
       }
     }
     return M;
   }

   /** Calcul des coefficients moyens sous l'hypothèse d'indépendance et
application d'une fonction linéaire au rapport 'valeur réelle'/'valeur
moyenne'
    @return M, matrice avec les nouveaux coefficients
    */
```

```
  public Matrice chi2Func() {
    Matrice M=new Matrice(m,n);
    double [][] X=M.getArray();
    double ni;
    double nj;
    double mij;
    double nij;
    double nT=sumTot();
    for (int i=0 ; i<m ; i++) {
      System.out.println("Ligne "+i);
      ni=sumLgn(i);
      for (int j=0 ; j<n ; j++) {
        nj=sumCol(j);
        if (nT!=0) {
          mij=ni*nj/nT;
          if (mij!=0) {
            nij=A[i][j];
            if (nij>=2*mij) {
              X[i][j]=1;
            }
            else if (nij<2*mij && nij>0) {
              X[i][j]=nij/(2*mij);
            }
          }
        }
      }
    }
    return M;
  }

  /** Applique la transformation tf-idf et retourne une nouvelle matrice
   */

  public Matrice tfidf() {
    int nbCoefNonNuls;
    Matrice Mat=new Matrice(m,n);
    double [][] X=Mat.getArray();
    double idf,coeff;

    for (int i=0 ; i<m;i++ ){
      nbCoefNonNuls = 0;

      for  (int j=0;j<n;j++){
        if (A[i][j]!=0) nbCoefNonNuls++;
      }

      idf = Math.log((double)n/(double)nbCoefNonNuls);

      for  (int j=0;j<n;j++){
        coeff = A[i][j];
        if (coeff != 0) {
          X[i][j]=coeff*idf;
        }
      }
    }
    return Mat;
  }

  /** Matrice des cooccurrences obtenue par produit de la matrice et de sa
transposée
   */
```

```
   public Matrice getMatriceCooc() {
     System.out.println("aaa");
     Matrice Mat = new Matrice(m,m);
     double [][] C = Mat.getArray();
     System.out.println("bbb");
     double s=0;
     double d=0;
     for (int j=0 ; j<m ; j++) {
       d=0;
       for (int k=0 ; k<n ; k++) {

         d+=A[j][k]*A[j][k];
       }
       C[j][j]=d;
       for (int i=0 ; i<m && i<j ; i++) {
         s=0;
         for (int k=0 ; k<n ; k++) {
           s+=A[i][k]*A[j][k];
         }
         C[i][j]=s;
         C[j][i]=s;
       }
       System.out.println("Matrice de cooccurrences : ligne "+j+"
calculée");
     }

     return Mat;
   }

   /** Applique une transformation de type LSA et retourne une nouvelle
matrice
    @param nbVP  nombre de valeurs singulieres conservees
    */
   public Matrice pseudoLSA(int nbVP) {
     System.out.println("Recours à SingularValueDecomposition,
patience...");
     SingularValueDecomposition2 SVD = new
SingularValueDecomposition2(this);
     System.out.println("Recuperaiton des valeurs singulieres...");
     double [] valSing = SVD.getSingularValues();
     int lgVS = valSing.length;
     if (nbVP>=lgVS) {
       return this;
     }
     else {
       double[][] D = new double [n][n];
       System.out.println("Generation de la matrice diagonale tronquee...");
       for (int i=0 ; i<nbVP ; i++) {
         D[i][i]=valSing[i];
       }
       Matrice Dmoins = new Matrice(D);
       System.out.println("Produit UD'V en cours...");
       return SVD.getU().times(Dmoins.timesTranspose(SVD.getV()));
     }
   }


   /** Matrice de taille (m,m) des cosinus fait entre tous les couples de
vecteurs-lignes (i,j)
    @return  M, matrice des cosinus
    */
```

```
public Matrice cosinus() {
  Matrice M=new Matrice(m,m);
  double [][] MA=M.getArray();


  double norm2Li;
  double norm2Lj;
  double produitScalaire;
  for (int i=0 ; i<m ; i++) {
    System.out.println("Cosinus ligne "+i);
    norm2Li=0;
    for (int j=0 ; j<n;j++) {
      norm2Li+=A[i][j]*A[i][j];
    }

    if (norm2Li==0) {
      MA[i][i]=0;
      for (int j=0 ; j<m && j<i; j++) {
        MA[i][j]=0;
        MA[j][i]=0;
      }
    }
    else {
      norm2Li=Math.sqrt(norm2Li);
      MA[i][i]=1;

      for (int j=0 ; j<m && j<i ; j++){

        norm2Lj=0;
        for (int k =0 ; k<n ; k++) {
          norm2Lj+=A[j][k]*A[j][k];
        }

        if (norm2Lj == 0) {
          MA[i][j]=0;
          MA[j][i]=0;
        }
        else {
          norm2Lj=Math.sqrt(norm2Lj);
          produitScalaire=0;
          for (int k=0;k<n;k++) {
            produitScalaire +=A[i][k]*A[j][k];
          }
          double cos = produitScalaire/(norm2Li*norm2Lj);
          MA[i][j]=cos;
          MA[j][i]=cos;
        }
      }
    }
  }
  return M;
}


/** Methode toString pour visualiser la matrice
 @return s, visualisation sous forme de tableau de la matrice
 */
public String toString() {
String s="";
for (int i=0 ; i<m ; i++) {
```

```
    for (int j=0 ; j<n ; j++) {
      s=s+A[i][j]+"  ";
    }
    s=s+"\n";
  }
  return s;
}
```

A2) Sémème de pollen, sable, éclat et or

Sémème de pollen

Le sémème affiché ci-dessous correspond aux informations délivrées en sortie de Sémy.

numéro	item
0	/4243/ : fluidement,ADV fluidique,ADJ fluidiste,NOM fluide,NOM fluidomètre,NOM fluidifiant,NOM fluidiforme,NOM fluidité,NOM suprafluidité,NOM fluidification,NOM superfluide,ADJ fluidifier,VERBE fluide,ADJ
1	être
2	/6635/ : saccageur,NOM saccageoter,VERBE saccageuse,NOM saccage,NOM saccagement,NOM saccager,VERBE sac,NOM
3	/2970/ : logement,NOM relogement,NOM loge,NOM logeur,NOM logeable,ADJ logeuse,NOM reloger,VERBE délogement,NOM logette,NOM loger,VERBE déloger,VERBE
4	poussière
5	/1101/ : preneur,NOM entr'ouvrir,VERBE reproductibilité,NOM décomposant,ADJ reprocheur,ADJ productif,ADJ composant,ADJ représenté,ADJ produire,VERBE improduit,ADJ incompréhensiblement,ADV entrouvrir,VERBE reproductivité,NOM mécomprendre,VERBE production,NOM rentré,NOM entrance,NOM représentation,NOM surprise,NOM appréhendement,NOM emprisonné,ADJ autoreproducteur,ADJ rentrayeur,NOM rentrant,NOM reprisage,NOM mécompréhension,NOM reproductif,ADJ incompréhensible,ADJ apprésentation,NOM prisonnier,NOM compréhension,NOM entr'ouvrement,NOM reproduire,VERBE plexus,NOM reproductrice,NOM pris,ADJ sentimentaliste,NOM sentimentalité,NOM surproduction,NOM entrouverture,NOM entreprise,NOM indécomposé,ADJ déprise,NOM reprographique,ADJ senti,NOM incomplexe,ADJ sentimentalisation,NOM irreprésentable,ADJ rentrante,NOM incompréhensif,ADJ reproche,NOM prisonnière,NOM rentrayeuse,NOM présent,ADJ prison,NOM représentée,NOM prise,NOM représenter,VERBE présenter,VERBE repriser,VERBE incompris,ADJ décomposer,VERBE reprise,NOM reprocher,VERBE présence,NOM procès,NOM compréhensible,ADJ représentante,NOM preneuse,NOM reproductivement,ADV entrer,VERBE reprisable,ADJ rentrant,ADJ représentativité,NOM reprographier,VERBE prendre,VERBE rentrage,NOM indécomposable,ADJ coproduction,NOM déprendre,VERBE repriseur,ADJ reprochable,ADJ reprendre,VERBE imprenable,ADJ composante,NOM présentation,NOM reproductible,ADJ improductivement,FUNC repriseuse,NOM coproduire,VERBE improductif,ADJ preneur,ADJ producteur,NOM répréhension,NOM préhension,NOM représentable,ADJ senti,ADJ entrant,NOM comprendre,VERBE reproduction,NOM rentrure,NOM entreprendre,VERBE incompréhensibilité,NOM appréhension,NOM sentimentaliser,VERBE produit,ADJ surproduit,NOM improductivité,NOM sentiment,NOM complexe,ADJ représenté,NOM représentatif,ADJ sentimental,ADJ prisonnier,ADJ produit,NOM reproducteur,ADJ intercompréhension,NOM sentimentalisme,NOM entrée,NOM rentrée,NOM sentir,VERBE emprisonner,VERBE dissentiment,NOM surproduire,VERBE appréhender,VERBE rentrer,VERBE rentré,ADJ surreprésentation,NOM apprésenter,VERBE entrant,ADJ présent,NOM composant,NOM représentativement,FUNC reproducteur,NOM incompréhension,NOM décomposition,NOM entrante,NOM entrepreneur,NOM emprisonnement,NOM

	surprendre,VERBE appréhendé,ADJ représentant,NOM reprographie,NOM décomposable,ADJ
6	/6024/ : membraniforme,ADJ membraneux,ADJ membranule,NOM membrané,ADJ membrane,NOM
7	anthère
8	très
9	/1054/ : petite,NOM petitement,ADV rapetisser,VERBE rapetissage,NOM petitette,ADJ rapetissement,NOM petitesse,NOM petiot,ADJ petiote,NOM petiot,NOM petit,ADJ apetisser,VERBE petitounet,ADJ apetissement,NOM petit,NOM
10	féconder
11	/352/ : jauniot,ADJ jaunissure,NOM jaunisse,NOM jaunet,ADJ jaunissement,NOM jaunasse,ADJ jaunissant,ADJ jaune,NOM jaune,ADJ jaunissage,NOM jaunir,VERBE jaunet,NOM jaunâtre,ADJ
12	/2482/ : entretenu,ADJ entreteneuse,NOM entretenage,NOM rétentrice,NOM entretenir,VERBE soutien,NOM soutènement,NOM rétentionnaire,ADJ codétenu,NOM entretènement,NOM rétenteur,ADJ retenir,VERBE détenir,VERBE soutenance,NOM codétenue,NOM contenir,VERBE soutenir,VERBE rétention,NOM détention,NOM rétenteur,NOM rétentionniste,NOM entreteneur,NOM retenu,ADJ détenu,NOM contention,NOM tenir,VERBE rétentionnel,ADJ entretien,NOM retenue,NOM rétentionnaire,NOM tenue,NOM
13	/8765/ : utriculaire,ADJ utriculeux,ADJ utricule,NOM
14	/107/ : agrainage,NOM grainetier,NOM grainetière,NOM agrainer,VERBE grainé,ADJ graine,NOM grainier,NOM grain,NOM agrain,NOM grainière,NOM graineterie,NOM grainasse,NOM
15	fin
16	généralement
17	/756/ : microcristal,NOM microscopiste,NOM microscopique,ADJ microphyte,NOM microbiologique,ADJ microscopie,NOM micrococcus,NOM microchirurgical,ADJ microchirurgie,NOM microbiologiste,NOM microzoaire,NOM microflore,NOM microbicide,ADJ micromètre,NOM microformes,NOM microdissection,NOM microlithe,NOM micrologique,ADJ microscopiquement,FUNC micrométriquement,FUNC amicrobien,ADJ micrographie,NOM microlite,NOM microfaune,NOM microtome,NOM micrologie,NOM micrométrie,NOM microfossile,NOM microscope,NOM microstructure,NOM inframicroscopique,ADJ micromanipulation,NOM microbien,ADJ microbiologie,NOM micrographique,ADJ microbicide,NOM microbisme,NOM microbiologiste,ADJ monomicrobien,ADJ micromanipulateur,NOM micrographe,NOM microorganisme,NOM ultramicroscope,NOM microbique,ADJ microcoque,NOM microbe,NOM ultramicroscopie,NOM micrométrique,ADJ

Sémème du mot sable

numéro	item
0	/2507/ : silicifié,ADJ silicotique,ADJ silicoformique,ADJ silicocyanhydrique,ADJ silicose,NOM silicocalcium,NOM silicium,NOM silicaté,ADJ silicate,NOM silicié,ADJ siliconage,NOM silicique,ADJ silicatiser,VERBE silicone,NOM silicosé,NOM silicocalcaire,ADJ silicater,VERBE silicomanganèse,NOM silicosée,NOM silicométhane,NOM silicosé,ADJ silicatage,NOM silicole,ADJ silicatisation,NOM silicogel,NOM siliceux,ADJ trisilicique,ADJ siliconer,VERBE silicicoleou,ADJ siliconé,ADJ silicochloroforme,NOM silicification,NOM silice,NOM silicatation,NOM siliciure,NOM
1	concrétion
2	/751/ : meublé,ADJ meuble,NOM démeublé,ADJ ameubler,VERBE meuble,ADJ meublé,NOM meublable,ADJ immeuble,ADJ ameublir,VERBE meublement,NOM remeubler,VERBE meublier,NOM ameublissement,NOM meublant,ADJ remeublement,NOM démeublement,NOM ameubli,ADJ ameublement,NOM meublage,NOM meubler,VERBE immeuble,NOM démeubler,VERBE
3	/533/ : transformer,VERBE formaliserse,VERBE fondant,ADJ profonde,NOM formaliste,ADJ profond,NOM profondeur,NOM néoformation,NOM formolage,NOM formuler,VERBE refondage,NOM préforme,NOM cofondatrice,NOM fortiori,FUNC forte,ADV fonder,VERBE biforme,ADJ reforming,NOM réformette,NOM préformant,ADJ fusionisme,NOM fondamentaliste,ADJ déformé,ADJ formalisé,ADJ préformé,ADJ formaliser,VERBE fusionner,VERBE forme,NOM forte,ADJ formulaire,NOM efforcement,NOM informatique,ADJ fondage,NOM formateur,NOM conformer,VERBE fuser,VERBE formulique,ADJ fondre,VERBE déformable,ADJ formulation,NOM préformage,NOM fondement,NOM fonderie,NOM informatrice,NOM format,NOM déformer,VERBE informaticien,NOM formateur,ADJ fondateur,NOM réformer,VERBE approfondi,ADJ approfondisseur,NOM fondée,NOM fondamentalité,NOM réformisme,NOM fusionnement,NOM réformiste,NOM déformation,NOM formeur,NOM formier,NOM formolateur,NOM surinformation,NOM périinformatique,NOM déformateur,ADJ fondu,NOM fusionniste,ADJ informatisation,NOM réformiste,ADJ informé,NOM transformation,NOM réformé,NOM cofondateur,NOM conforme,ADJ formellement,ADV informité,NOM formolisation,NOM formol,NOM formiate,NOM informatif,ADJ informaticienne,NOM téléinformatique,NOM déformant,ADJ informationnel,ADJ confusionnisme,NOM effondrer,VERBE parfondre,VERBE information,NOM préformation,NOM réformé,ADJ réformée,NOM informel,ADJ informé,ADJ fusionnage,NOM approfondir,VERBE refonte,NOM informatiser,VERBE infondé,ADJ informant,ADJ refusion,NOM confondre,VERBE formant,NOM refondre,VERBE info,NOM fondé,ADJ réformation,NOM réformateur,NOM formage,NOM approfondissant,ADJ formatrice,NOM méforme,NOM informateur,NOM formant,ADJ réformatrice,NOM formolé,ADJ effondrement,NOM formique,ADJ formalisable,ADJ reformage,NOM formaliste,NOM approfondissement,NOM réforme,NOM informer,VERBE conformateur,NOM formulable,ADJ fortiori a,FUNC fondamentaliste,NOM reformation,NOM fondue,NOM fusion,NOM reformulation,NOM confusion,NOM fondatrice,NOM informulable,ADJ formalité,NOM fondamental,ADJ fondé,NOM formoler,VERBE effondrilles,NOM efforcer,VERBE fondoir,NOM profond,ADJ irréformable,ADJ effort,NOM formalisant,ADJ fortement,ADV informatisé,ADJ fort,ADJ formalisme,NOM préformer,VERBE irréformabilité,NOM fondamentalement,ADV reformuler,VERBE reformer,VERBE formel,ADJ superforme,NOM fonte,NOM informulé,ADJ fusage,NOM infondre,VERBE uniformément,ADV former,VERBE

	réformateur,ADJ conformation,NOM indéformabilité,NOM fondeur,NOM informatique,NOM fond,NOM formatif,ADJ réformage,NOM fondant,NOM formation,NOM fondation,NOM formalisation,NOM fondu,ADJ fusionnisme,NOM indéformable,ADJ formule,NOM profondément,ADV informateur,ADJ conformément,ADV informe,ADJ
4	/8812/ : vastité,NOM vaste,ADJ vastitude,NOM
5	étendue
6	pulvérulent
7	/233/ : accordement,NOM solidien,ADJ accolade,NOM incomplet,NOM couple,NOM accordable,ADJ inaccord,NOM solide,ADJ pulsionnel,ADJ structuraliste,NOM cordonnage,NOM accourci,NOM courson,NOM solidement,ADV complice,ADJ solidité,NOM cordonnet,NOM malcommode,ADJ restructurer,VERBE recomposable,ADJ accommodement,NOM constitution,NOM cordeler,VERBE reconstructeur,ADJ mercerie,NOM recordage,NOM complètement,NOM accommodat,NOM cordier,ADJ structuration,NOM corder,VERBE compulsion,NOM accouplage,NOM raccompagnement,NOM restructuration,NOM anticonstitutionnel,ADJ mercurochrome,NOM commode,ADJ accommodation,NOM corderie,NOM réaccorder,VERBE mercure,NOM court-circuitage,NOM cordé,ADJ structuralisme,NOM reconstituant,ADJ commercer,VERBE accord,NOM infrastructure,NOM coursonne,NOM commerce,NOM mercantilisme,NOM décordage,NOM incommode,ADJ courtaud,NOM complice,NOM surcomposé,ADJ reconstituable,ADJ accordailles,NOM compléter,VERBE accourcissement,NOM inconstitutionnel,ADJ cordelle,NOM inaccompli,ADJ compagnon,NOM cordier,NOM compagne,NOM accommodant,ADJ cordon,NOM accoler,VERBE reconstituer,VERBE consolider,VERBE reconstruire,VERBE surcomposer,VERBE inaccompli,NOM accoupler,VERBE mercantilisation,NOM décordement,NOM mercantiliste,ADJ constituer,VERBE composition,NOM mercier,NOM accorder,VERBE accompli,ADJ découpler,VERBE accouplé,ADJ structuralement,FUNC pulsion,NOM accomplissement,NOM accordant,ADJ accommodateur,ADJ accompagnateur,NOM mercuriel,ADJ incomplète,NOM cordage,NOM structurellement,ADV cordophone,NOM cordière,NOM compulser,VERBE incomplet,ADJ accolement,NOM accommodeuse,NOM accourci,ADJ incommodité,NOM découpler,NOM encordage,NOM accompagner,VERBE cordelier,NOM astructurel,ADJ inconfort,NOM inconvénient,NOM incommodément,ADV reconstructeur,NOM composer,VERBE reconstructrice,NOM confortement,NOM courtauder,VERBE cordelé,ADJ complétion,NOM inconstitutionnellement,ADV recomplétement,NOM découple,NOM accordeur,NOM mercantiliser,VERBE accourcie,NOM accommodante,NOM couplage,NOM structure,NOM accommodant,NOM accolader,VERBE cordonné,ADJ structurant,ADJ recomposer,VERBE ultrastructure,NOM cordelette,NOM accolure,NOM complet,ADJ compulsation,NOM monocorde,ADJ accordoir,NOM accommoder,VERBE composé,ADJ accordage,NOM accordant,NOM inaccordable,ADJ encorder,VERBE inaccommodable,ADJ structurable,ADJ accommodage,NOM décorder,VERBE incomplètement,ADV incomplétude,NOM accolage,NOM monocorde,NOM rétropulsion,NOM raccompagnade,NOM commodité,NOM accouplement,NOM courtaude,NOM reconstitution,NOM consolidement,NOM solidifier,VERBE mercanti,NOM cordeau,NOM court,ADJ courson,ADJ structuraliste,ADJ réaccord,NOM cordelière,NOM structurer,VERBE couplé,ADJ encordement,NOM accompagnateur,ADJ surstructure,NOM accompagnement,NOM cordonner,VERBE accomplir,VERBE accort,ADJ structurel,ADJ consolidation,NOM réaccordage,NOM confort,NOM accommodatif,ADJ inconstitutionnalité,NOM mercureux,ADJ recompléter,VERBE anticonstitutionnellement,ADV consolidé,ADJ mercantiliste,NOM courtement,ADV déconstruction,NOM accomplisseur,NOM couplement,NOM construire,VERBE pulser,VERBE reconstruction,NOM accommodeur,NOM accordance,NOM

	accommodable,ADJ mercière,NOM découplé,ADJ accolerie,NOM complicité,NOM coupleur,NOM construction,NOM accourcir,VERBE mercantile,ADJ cordée,NOM courtilière,NOM recorder,VERBE accompagnant,ADJ découplage,NOM conforter,VERBE courtcircuiter,VERBE solidification,NOM corde,NOM complètement,ADV découplement,NOM structuré,ADJ commercial,ADJ accompagnatrice,NOM courtaud,ADJ déconstruire,VERBE précomplètement,NOM pulseur,NOM paracommercial,ADJ réaccordement,NOM raccompagner,VERBE recomposition,NOM structural,ADJ coupler,VERBE surcomposition,NOM
8	/4298/ : fragmentairement,ADV fragmentarité,NOM fragmentaire,ADJ fragment,NOM fragmentarisme,NOM fragmenter,VERBE fragmenté,ADJ fragmentier,NOM fragmentation,NOM
9	sédimentaire
10	/2621/ : coule,NOM couleur,NOM
11	émail
12	/5364/ : inutilité,NOM inutiliser,VERBE inutilement,ADV utilitariste,NOM utilité,NOM réutiliser,VERBE utilitairement,FUNC utilisatrice,NOM utilisateur,NOM inutile,ADJ utilitariste,ADJ utilitarisme,NOM utilitaire,ADJ utilisable,ADJ inutilisable,ADJ utile,NOM utilisation,NOM utilisateur,ADJ inutilisé,ADJ utiliser,VERBE utilement,ADV inutilisation,NOM réutilisation,NOM utile,ADJ
13	/61/ : paralléliseur,NOM antiparasite,NOM dépareillé,ADJ apparaître,VERBE parer,VERBE repassage,NOM parascève,NOM passepied,NOM reparaître,VERBE comparaître,VERBE préparage,NOM pareuse,NOM passerelle,NOM appareil,NOM impassable,ADJ apparition,NOM disparaître,VERBE antiparasite,ADJ passante,NOM apparaux,NOM passavant,NOM repasser,VERBE apparat,NOM repasseur,NOM passéifier,VERBE pareur,NOM appareilleur,NOM passure,NOM paresthésie,NOM dépasser,VERBE apparution,NOM passerine,NOM passager,VERBE parasite,NOM passegrand,ADJ passéisme,NOM passéification,NOM repasseuse,NOM parader,VERBE antiparasitaire,NOM déparer,VERBE dépassant,NOM surpasser,VERBE paraître,VERBE passagère,NOM parage,NOM surpassement,NOM pareil,ADJ insurpassé,ADJ passoire,NOM dépassante,NOM appareiller,VERBE passe,NOM passeur,NOM passation,NOM passéiste,ADJ passade,NOM impréparation,NOM apparent,NOM antiparasitaire,ADJ passegrande,ADJ passement,NOM passériformes,NOM repasse,NOM passager,NOM préparer,VERBE réapparaître,VERBE impréparé,ADJ appareillement,NOM apparoir,VERBE passablement,ADV appareillé,ADJ reparution,NOM passager,ADJ passé,NOM comparoir,VERBE préparation,NOM indépassable,ADJ paradigme,NOM inapparent,ADJ passementer,VERBE parade,NOM passage,NOM dépassement,NOM parution,NOM indépassé,ADJ appareillage,NOM insurpassable,ADJ réapparition,NOM apparaissance,NOM apparence,NOM passée,NOM passant,NOM pas,NOM disparition,NOM passéiste,NOM passette,NOM apparent,ADJ passé,ADJ préparatif,NOM dépassant,ADJ impasse,NOM passereau,NOM passeport,NOM parement,NOM passer,VERBE apparemment,ADV passant,ADJ apparente,NOM comparution,NOM passable,ADJ
14	désertique
15	/778/ : amortissable,ADJ amorti,ADJ amortissage,NOM amortie,NOM amorti,NOM amortir,VERBE mortier,NOM amortissement,NOM
16	/1054/ : petite,NOM petitement,ADV rapetisser,VERBE rapetissage,NOM petitette,ADJ rapetissement,NOM petitesse,NOM petiot,ADJ petiote,NOM petiot,NOM petit,ADJ apetisser,VERBE petitounet,ADJ apetissement,NOM petit,NOM
17	/3021/ : nombré,ADJ dénombrement,NOM indénombrable,ADJ nombrant,ADJ dénombrable,ADJ dénombrer,VERBE nombrage,NOM innombrable,ADJ nombreux,ADJ nombreusement,FUNC innombrablement,ADV nombrer,VERBE numératif,ADJ

	nombrable,ADJ nombre,NOM innombrabilité,NOM numération,NOM
18	/2401/ : claironnement,NOM clairet,ADJ clairière,NOM clair,ADJ clairvoyance,NOM inclairvoyance,NOM clair,NOM clairvoyant,ADJ clairon,NOM claire,NOM claironnée,NOM inclairvoyant,ADJ claironnant,ADJ clairement,ADV claironner,VERBE clairette,NOM clairsemé,ADJ clairet,NOM clairsemer,VERBE claironné,ADJ
19	/1364/ : noircissage,NOM noirâtre,ADJ noirot,NOM noircisseur,NOM noircissement,NOM noiraud,ADJ noireau,NOM noire,NOM noircissant,ADJ noir,ADJ noirien,NOM noirouffe,ADJ noirement,FUNC noirceur,NOM noircisseur,ADJ noircisseuse,NOM noircir,VERBE noircissure,NOM noir,NOM noirin,NOM
20	/3707/ : sablonnière,NOM sablonner,VERBE sablerie,NOM sablonneux,ADJ sabler,VERBE ensablage,NOM ensablement,NOM sablage,NOM ensabler,VERBE sable,NOM sableux,ADJ sablonnier,NOM sablier,NOM sablon,NOM sableur,NOM sablé,ADJ ensablé,ADJ sablière,NOM sableuse,NOM
21	désagrégation
22	/2169/ : cavicole,NOM concavité,NOM cavité,NOM cavicole,ADJ supercavitant,ADJ cavitaire,ADJ cavitation,NOM
23	/8754/ : urinement,NOM uriner,VERBE urinaire,ADJ urinifère,ADJ urinal,NOM urine,NOM urination,NOM urineux,ADJ urinage,NOM urinoir,NOM
24	/291/ : survenir,VERBE aventurer,VERBE survenue,NOM revendicateur,NOM aventureux,NOM aventureuse,NOM revertier,NOM aventure,NOM revenant,NOM advenant,ADJ aventurier,ADJ aventureusement,ADV avenir,VERBE survenance,NOM advenir,VERBE prévenir,VERBE aventurière,NOM revenue,NOM revendicatif,ADJ aventurier,NOM revenu,ADJ souvenirse,VERBE venue,NOM revenant,ADJ aventureux,ADJ avènement,NOM aventuré,ADJ revendicateur,ADJ souvenance,NOM avenir,NOM souvenir,NOM souvenir,VERBE prévention,NOM revendiquer,VERBE mésavenant,ADJ revendicatrice,NOM parvenir,VERBE revendication,NOM revendicant,ADJ revenu,NOM revenante,NOM provenir,VERBE aventurisme,NOM advènement,NOM revenir,VERBE venir,VERBE
25	/3081/ : rochasse,NOM rocher,NOM rocheux,ADJ dérochage,NOM dérocher,VERBE roche,NOM enrochement,NOM dérochement,NOM enrocher,VERBE
26	/996/ : dénaturant,NOM nature,NOM naturalisme,NOM antinaturel,ADJ supernaturalisme,NOM dénaturalisation,NOM dénaturé,ADJ extranaturel,ADJ naturel,NOM antinaturalisme,NOM supernaturel,ADJ naturant,ADJ naturalisé,NOM naturaliser,VERBE naturellement,ADV naturaliste,ADJ surnaturel,ADJ naturiste,ADJ dénaturement,NOM naturel,ADJ dénaturation,NOM naturé,ADJ supernaturaliste,ADJ connaturel,ADJ dénaturer,VERBE dénaturant,ADJ naturalisé,ADJ naturalisée,NOM naturalité,NOM surnature,NOM naturalisation,NOM naturaliste,NOM dénaturaliser,VERBE naturisme,NOM naturelle,NOM
27	/107/ : agrainage,NOM grainetier,NOM grainetière,NOM agrainer,VERBE grainé,ADJ graine,NOM grainier,NOM grain,NOM agrain,NOM grainière,NOM graineterie,NOM grainasse,NOM
28	/79/ : hémiorganisme,NOM organopathie,NOM inorganisation,NOM organiser,VERBE organicienne,NOM organe,NOM organosol,NOM réorganiser,VERBE organogenèse,NOM inorganique,ADJ organiquement,ADV réorganisatrice,NOM organotaxie,NOM organisateur,NOM organicité,NOM organisation,NOM organostannique,ADJ organogel,NOM anorganique,ADJ orgasme,NOM orgastique,ADJ organogénèse,NOM réorganisateur,NOM organisatrice,NOM inorganisé,ADJ organométallique,ADJ hyperorganique,NOM réorganisation,NOM organisationnel,ADJ organicisme,NOM organisable,ADJ organisme,NOM réorganisateur,ADJ organismique,ADJ organicien,ADJ inorganisable,ADJ organique,ADJ organogène,ADJ

	organographie,NOM organothérapie,NOM organodynamisme,NOM organisant,ADJ organicien,NOM hyperorganisme,NOM organiciste,ADJ organisé,ADJ
29	notamment
30	/9235/ : confection,NOM confectionnement,NOM confectionner,VERBE
31	beige
32	/2546/ : substance,NOM consubstantiel,ADJ insubstance,NOM insubstantiel,ADJ consubstantialité,NOM insubstantialité,NOM transsubstantiation,NOM substantiel,ADJ transsubstantier,VERBE
33	/3268/ : diversifier,VERBE divers,ADJ diversification,NOM diversité,NOM diversement,ADV

Sémème du mot éclat

numéro	item
0	/3687/ : illuminé,NOM enluminage,NOM lumineux,ADJ luminifère,ADJ luminescent,ADJ enlumineuse,NOM illuminatrice,NOM luministe,ADJ luministe,NOM illuminer,VERBE illuminé,ADJ illuminant,ADJ lumière,NOM luminariste,NOM illuminée,NOM enluminure,NOM luminance,NOM illuminateur,NOM illumination,NOM illuministe,ADJ lumineusement,ADV lumination,NOM luminosité,NOM enluminer,VERBE luminaire,NOM superluminique,ADJ illuministe,NOM illuminisme,NOM illuminateur,ADJ luminescence,NOM illumineur,NOM enlumineur,NOM
1	bruyant
2	/524/ : munitionnaire,NOM munition,NOM munitionner,VERBE démuni,ADJ démunition,NOM munir,VERBE démunir,VERBE
3	/4974/ : manifestation,NOM monomanie,NOM immaniable,ADJ manichéen,ADJ remaniable,ADJ manifestant,NOM manifestement,ADV manieuse,NOM manifeste,ADJ hypomanie,NOM manipulateur,NOM maniériste,ADJ manipulatrice,NOM maniaque,NOM manifestable,ADJ maniérer,VERBE maniement,NOM maniérisme,NOM manipuler,VERBE manière,NOM manichéen,NOM maniériste,NOM maniaque,ADJ manichéenne,NOM manieur,NOM manie,NOM maniéré,ADJ maniériser,VERBE manipulation,NOM maniaquement,FUNC télémanipulateur,NOM manichéisme,NOM maniage,NOM manipulable,ADJ maniaquerie,NOM manier,VERBE remaniement,NOM manifestante,NOM mani,UNDEF manifester,VERBE remanier,VERBE
4	intensité
5	réfléchir
6	violent
7	/322/ : affinitaire,ADJ affinant,ADJ affin,NOM affineur,NOM affinostat,NOM affilier,VERBE affiliation,NOM affine,NOM affiner,VERBE affineuse,NOM affinée,NOM affinoir,NOM affinement,NOM affinerie,NOM finesse,NOM affin,ADJ confinement,NOM affinitaire,NOM affineur,ADJ affiné,ADJ confinage,NOM affination,NOM affinité,NOM confiner,VERBE superfinement,FUNC affiné,NOM affinage,NOM
8	/2585/ : intrus,NOM incorporellement,FUNC injecteur,ADJ corporéité,NOM réinjecter,VERBE introduire,VERBE corporisation,NOM réintroduire,VERBE corporatiste,NOM surcorps,NOM intégrer,VERBE intégralisme,NOM intégration,NOM corpusculaire,ADJ intruse,NOM incorporable,ADJ introductrice,NOM incorporer,VERBE introniser,VERBE intégral,ADJ corporellement,ADV réincorporer,VERBE réintroduction,NOM incorporel,ADJ corporatisme,NOM corpulence,NOM corporatiste,ADJ corporiser,VERBE incorporation,NOM intégrabilité,NOM injecteur,NOM réintégration,NOM intégratif,ADJ réintégrer,VERBE corporification,NOM intégrationniste,NOM réinjection,NOM incorporant,ADJ intégrateur,NOM réintégrable,ADJ incorporalité,NOM introducteur,NOM corporatif,ADJ corporel,ADJ intrusion,NOM introductoire,ADJ injecté,ADJ intégré,ADJ intromission,NOM corporation,NOM réintégrande,NOM injectable,ADJ corporativement,ADV incorporéité,NOM intégrationniste,ADJ corporifier,VERBE corps,NOM réincorporation,NOM corporence,NOM intégralement,ADV introduction,NOM corporalité,NOM intronisation,NOM intégrale,NOM corpuscule,NOM intégrable,ADJ corpulent,ADJ injection,NOM injecter,VERBE intrusif,ADJ
9	/6302/ : naissant,ADJ naistre,null naître,VERBE renaissant,ADJ renaître,VERBE

	naissance,NOM renaissance,NOM
10	/8619/ : tonnerre,NOM tonnant,ADJ tonner,VERBE
11	détaché
12	/8645/ : touffe,NOM touffer se,FUNC touffer,FUNC touffette,NOM touffu,ADJ
13	/4298/ : fragmentairement,ADV fragmentarité,NOM fragmentaire,ADJ fragment,NOM fragmentarisme,NOM fragmenter,VERBE fragmenté,ADJ fragmentier,NOM fragmentation,NOM
14	surtout
15	/6427/ : renouvelante,NOM nouveau,ADJ renouvelant,ADJ nouveauté,NOM renouveau,NOM nouvel,ADJ renouvelant,NOM nouvelleté,NOM renouveler,VERBE renouvellement,NOM renouvelé,ADJ renouvelable,ADJ
16	/1862/ : brisée,NOM brisis,NOM brisement,NOM brisé,ADJ brisable,ADJ brisure,NOM débris,NOM brisant,NOM bris,NOM briser,VERBE brisant,ADJ brisage,NOM briseur,NOM imbrisable,ADJ brisoir,NOM
17	esprit
18	caractère
19	source
20	éclatement
21	ton
22	capacité
23	/818/ : analyticité,NOM analytique,NOM analogue,NOM analogique,ADJ analyste,NOM analogicité,NOM analytique,ADJ analogiquement,ADV analogue,ADJ inanalysablement,FUNC inanalysé,ADJ inanalysable,ADJ analyse,NOM analogiste,NOM analogie,NOM analyseur,NOM analysable,ADJ analyser,VERBE analytiquement,ADV
24	déchirement
25	mérite
26	vivacité
27	/1101/ : preneur,NOM entr'ouvrir,VERBE reproductibilité,NOM décomposant,ADJ reprocheur,ADJ productif,ADJ composant,ADJ représenté,ADJ produire,VERBE improduit,ADJ incompréhensiblement,ADV entrouvrir,VERBE reproductivité,NOM mécomprendre,VERBE production,NOM rentré,NOM entrance,NOM représentation,NOM surprise,NOM appréhendement,NOM emprisonné,ADJ autoreproducteur,ADJ rentrayeur,NOM rentrant,NOM reprisage,NOM mécompréhension,NOM reproductif,ADJ incompréhensible,ADJ apprésentation,NOM prisonnier,NOM compréhension,NOM entr'ouvrement,NOM reproduire,VERBE plexus,NOM reproductrice,NOM pris,ADJ sentimentaliste,NOM sentimentalité,NOM surproduction,NOM entrouverture,NOM entreprise,NOM indécomposé,ADJ déprise,NOM reprographique,ADJ senti,NOM incomplexe,ADJ sentimentalisation,NOM irreprésentable,ADJ rentrante,NOM incompréhensif,ADJ reproche,NOM prisonnière,NOM rentrayeuse,NOM présent,ADJ prison,NOM représentée,NOM prise,NOM représenter,VERBE présenter,VERBE repriser,VERBE incompris,ADJ décomposer,VERBE reprise,NOM reprocher,VERBE présence,NOM procès,NOM compréhensible,ADJ représentante,NOM preneuse,NOM reproductivement,ADV entrer,VERBE reprisable,ADJ rentrant,ADJ représentativité,NOM reprographier,VERBE prendre,VERBE rentrage,NOM indécomposable,ADJ coproduction,NOM déprendre,VERBE repriseur,ADJ reprochable,ADJ reprendre,VERBE imprenable,ADJ composante,NOM présentation,NOM reproductible,ADJ improductivement,FUNC repriseuse,NOM coproduire,VERBE improductif,ADJ preneur,ADJ producteur,NOM

	répréhension,NOM préhension,NOM représentable,ADJ senti,ADJ entrant,NOM comprendre,VERBE reproduction,NOM rentrure,NOM entreprendre,VERBE incompréhensibilité,NOM appréhension,NOM sentimentaliser,VERBE produit,ADJ surproduit,NOM improductivité,NOM sentiment,NOM complexe,ADJ représenté,NOM représentatif,ADJ sentimental,ADJ prisonnier,ADJ produit,NOM reproducteur,ADJ intercompréhension,NOM sentimentalisme,NOM entrée,NOM rentrée,NOM sentir,VERBE emprisonner,VERBE dissentiment,NOM surproduire,VERBE appréhender,VERBE rentrer,VERBE rentré,ADJ surreprésentation,NOM apprésenter,VERBE entrant,ADJ présent,NOM composant,NOM représentativement,FUNC reproducteur,NOM incompréhension,NOM décomposition,NOM entrante,NOM entrepreneur,NOM emprisonnement,NOM surprendre,VERBE appréhendé,ADJ représentant,NOM reprographie,NOM décomposable,ADJ
28	/36/ : interposition,NOM interjecter,VERBE abréaction,NOM soulignage,NOM intercalage,NOM intersecter,VERBE port,NOM psychodramatique,ADJ inintelligibilité,NOM lignage,NOM aligneur,NOM transportable,ADJ psychotrope,NOM interruptif,ADJ imposeur,NOM interstice,NOM positionner,VERBE médiatisation,NOM réactiver,VERBE psychométrique,ADJ positivement,ADV psychopathologique,ADJ interfoliage,NOM proposition,NOM terminologie,NOM tendanciellement,FUNC défini,ADJ décidément,ADV terminateur,ADJ psychanalyser,VERBE abrupt,ADJ portière,NOM ventriculostomie,NOM télédétection,NOM inintelligent,ADJ transport,NOM pensionnement,NOM indispensable,ADJ entendement,NOM importable,ADJ tenter,VERBE intempestivement,ADV portement,NOM portefeuille,NOM possessif,ADJ possibiliser,VERBE interférométrique,ADJ psychologue,ADJ intelligiblement,ADV aéroport,NOM psycholinguistique,NOM activation,NOM rapport,NOM surpousse,NOM réactivation,NOM intense,ADJ pensionné,NOM abrupt,NOM actif,ADJ abréagir,VERBE exposante,NOM exigible,ADJ lignée,NOM intellectualité,NOM inactivation,NOM rétroactivement,ADV psychiatrique,ADJ médiateur,NOM interruptrice,NOM portatif,NOM souligner,VERBE réactionnaire,ADJ terminaison,NOM déposement,NOM interloquer,VERBE poseur,ADJ psychonévrotique,ADJ reposer,VERBE exporter,VERBE pensionnat,NOM lignager,NOM interpolateur,ADJ indéterminisme,NOM intelligemment,ADV portioncule,NOM inentendu,ADJ porte,NOM psychophysiologiste,NOM hypotenseur,ADJ intensif,NOM enligner,VERBE déposition,NOM repose,NOM approchement,NOM prédétermination,NOM actioniste,NOM sectionnement,NOM tendu,ADJ surcompensé,ADJ psychanalyste,NOM psyché,NOM interligné,ADJ psychopédagogique,ADJ indéfiniment,ADV rétroagissant,ADJ reporteur,NOM intransportable,ADJ surtension,NOM psychothérapie,NOM hypertendu,ADJ effranger,VERBE déterminabilité,NOM repossession,NOM poussée,NOM pension,NOM possibilité,NOM ventriculite,NOM dépouillement,NOM tendue,NOM actinique,ADJ psychographique,ADJ rapporteur,ADJ psychotrope,ADJ possessivité,NOM intermédiate,NOM transporter,VERBE interpolation,NOM atermoyeur,NOM repoussement,NOM psychologiste,ADJ psychisme,NOM aligné,ADJ positionnement,NOM interventionnisme,NOM comporter,VERBE impossibilité,NOM possessoire,NOM tente,NOM indisposé,ADJ surdétermination,NOM emporter,VERBE terminologique,ADJ psycholepsie,NOM pensionner,VERBE importer,VERBE psychogénèse,NOM malintention,NOM imposant,ADJ psychologisation,NOM déterministe,NOM interaction,NOM récompensant,ADJ tentatif,ADJ activeur,NOM alignement,NOM transposer,VERBE indisposition,NOM indéfini,NOM dépotoir,NOM terminateur,NOM appréciateur,NOM mésinterprétation,NOM terminisme,NOM interférométrie,NOM poussement,NOM psycholeptique,ADJ remporter,VERBE porteuse,NOM terminal,NOM suractif,ADJ rétroactes,NOM agissement,NOM indéterminabilité,NOM médiat,ADJ portager,VERBE intermission,NOM approchage,NOM indéterminément,FUNC réagir,VERBE inappréciation,NOM

	psychosomatique,ADJ exacteur,NOM transigeance,NOM tendu,NOM décider,VERBE ininterprété,ADJ intensif,ADJ tensionnement,NOM exposant,NOM intervenir,VERBE poussette,NOM intellectualisation,NOM alignée,NOM prédéterminer,VERBE hyperactivité,NOM approximatif,NOM entendre,VERBE interpsychologie,NOM indéterministe,ADJ exportatrice,NOM tentement,NOM tentelette,NOM inapprochable,ADJ interpolateur,NOM interventionniste,ADJ intelligence,NOM inapprécié,ADJ compensation,NOM suractivation,NOM actionné,ADJ intello,NOM inappréciablement,FUNC pensionné,ADJ médiumnité,NOM interprétatif,ADJ appréciation,NOM psychanalysé,ADJ psycholeptique,NOM portulan,NOM rapporter,VERBE psychanalysé,NOM interface,NOM inactiver,VERBE indéfinie,NOM repoussant,ADJ intellectuel,ADJ proportionner,VERBE malintentionné,ADJ reposée,NOM psychopédagogue,NOM récompense,NOM médiatisable,ADJ psychothérapeute,NOM porteur,ADJ interminable,ADJ déterminé,NOM approximer,VERBE abruptement,ADV impossible,NOM réactivité,NOM activisme,NOM session,NOM déterministe,ADJ imposition,NOM psychogénique,ADJ tendanciel,ADJ réacteur,NOM distension,NOM interférer,VERBE franger,VERBE appréciatif,ADJ portefeuilliste,NOM apporteur,NOM pensionnaire,NOM portière,ADJ tendage,NOM rompeur,ADJ indéfinitude,NOM intentionnaliser,VERBE psycholinguistique,ADJ portantine,NOM lignerolle,NOM intelligentiel,ADJ portefaix,NOM ligne,NOM positiviste,NOM exposer,VERBE compenser,VERBE actiniquement,ADV porteur,NOM ininterprétable,ADJ tentatrice,NOM possessionnel,ADJ réactivement,FUNC intelligentsia,NOM psychologique,ADJ récompenseur,NOM entendeur,NOM intensifier,VERBE porterie,NOM psychogenèse,NOM intenter,VERBE détendeur,NOM intensification,NOM compossibilité,NOM dispenser,VERBE porté,NOM préexponentiel,ADJ psychosomaticien,NOM psychique,ADJ psychopharmacologie,NOM activateur,ADJ interpolatrice,NOM possédante,NOM suractiver,VERBE dépositoire,NOM intensivement,ADV préhypertendu,ADJ psychodiagnostic,NOM imposance,NOM intercéder,VERBE psychiatre,NOM porté,ADJ rapportage,NOM psychotique,ADJ psychanalytique,ADJ positionniste,NOM indéfinissable,ADJ psychophysiologique,ADJ psychosomaticienne,NOM rompeuse,NOM antipsychiatrie,NOM réimposition,NOM poussage,NOM psychosexuel,ADJ prédéterminant,ADJ sectionneur,NOM interprétante,NOM psychométricien,NOM interprétable,ADJ dépens,NOM réactif,NOM actif,NOM lignard,NOM intervenant,NOM portion,NOM interlocutoire,ADJ reporter,NOM appréciable,ADJ psychologie,NOM impossiblement,FUNC activement,ADV proche,ADJ intercepteur,NOM intercepter,VERBE téléreporter,NOM prédéterminisme,NOM mésinterpréter,VERBE tensionner,VERBE supraventriculaire,ADJ rapporteuse,NOM préhypertendue,NOM psychographie,NOM actionnariat,NOM interruption,NOM exportation,NOM terminologue,NOM intelligible,ADJ interminablement,ADV interagir,VERBE interprétation,NOM rompement,NOM interpoler,VERBE psychothérapique,ADJ indéfini,ADJ rupturer,VERBE déposer,VERBE psychotique,NOM exponentiellement,ADV interminé,ADJ interlignage,NOM déterminable,ADJ médiateur,ADJ hypertension,NOM approcher,VERBE exiger,VERBE important,ADJ intempestivité,NOM imposer,VERBE détenteur,NOM frangette,NOM indéterminer,VERBE intermédine,NOM apprécié,ADJ inexigible,ADJ réactrice,NOM précieux,NOM réexporter,VERBE tendreté,NOM terme,NOM repousser,VERBE indécis,ADJ intentionnalité,NOM appréciatrice,NOM intensificateur,NOM déterminant,ADJ poussé,ADJ pensionnée,NOM dépost,NOM frange,NOM intermédiarité,NOM acter,VERBE intentionaliser,VERBE psychoneurologue,NOM disruptif,ADJ report,NOM portemanteau,NOM pousseuse,NOM rétroaction,NOM psychosocial,ADJ métapsychologie,NOM psychologiste,NOM inactif,ADJ interligneur,NOM tende de tranche,NOM reporter,VERBE transigement,NOM prépsychose,NOM inexigibilité,NOM actionnel,ADJ atermoiement,NOM téléreportage,NOM exportateur,NOM tension,NOM rapporteur,NOM activité,NOM exigence,NOM pousseur,NOM transaction,NOM distendre,VERBE inintelligence,NOM

	possédant,ADJ portier,NOM entestement,null psychomoteur,ADJ psychodrame,NOM intercaler,VERBE reposoir,NOM interjectif,ADJ lignomètre,NOM superposition,NOM intensément,ADV interligne,NOM enlignement,NOM surcompenser,VERBE intellectualisant,ADJ intermittemment,FUNC apposition,NOM positiver,VERBE interlinéaire,ADJ réactif,ADJ psychanalysée,NOM pousse,NOM réexportation,NOM apporteuse,NOM terminé,ADJ ventriculoscopie,NOM déterminant,NOM exportateur,ADJ coaction,NOM intercalement,NOM intentionalité,NOM portelet,NOM psychoprophylaxie,NOM poussoir,NOM frangère,NOM psychasthénie,NOM porte,ADJ hypotension,NOM hypotensif,ADJ psychogène,ADJ polypsychisme,NOM apposement,NOM superposable,ADJ impossible,ADJ décision,NOM interfolier,VERBE intelligent,ADJ interféromètre,NOM psychiatrisée,NOM hypertensif,ADJ entente,NOM interférentiel,ADJ interposé,ADJ psychogérontologue,NOM portoir,NOM interjectionnel,ADJ imposé,ADJ psychogénèse,NOM rétroagir,VERBE transposition,NOM indéterminé,NOM intentionniste,NOM ligner,VERBE psychasthénique,ADJ abréacteur,NOM disposition,NOM psychose,NOM détermination,NOM définisseur,NOM psychopédagogie,NOM ventriculogramme,NOM pose,NOM terminer,VERBE intermédiat,NOM repoussé,ADJ rétrospectivement,ADV activer,VERBE intermittent,ADJ impositionnaire,NOM intercesseur,NOM intentionnellement,ADV poseuse,NOM reposition,NOM tensioactif,ADJ psychobiologie,NOM tendre,VERBE tentative,NOM tensiomètre,NOM ventriculométrie,NOM effrangement,NOM rupture,NOM action,NOM psychomotricité,NOM déposante,NOM psychogénie,NOM lignerole,NOM réactimètre,NOM portant,NOM déterminante,NOM hypertendu,NOM psychonévrosé,NOM portionnette,NOM intello,ADJ tensionnage,NOM portée,NOM psychiatrisation,NOM malentendu,NOM réaligner,VERBE transportation,NOM médiatrice,NOM mésentente,NOM interprète,NOM interminis,FUNC interrompre,VERBE approximatif,ADJ psychiatrisé,NOM approche,NOM décompensation,NOM déterminer,VERBE exponentiel,ADJ psychopolynévrite,NOM possiblement,FUNC intellectualiste,ADJ psychologue,NOM rétroactif,ADJ intellection,NOM ventriculaire,ADJ dépôt,NOM psychomoral,ADJ terminal,ADJ transporter,NOM interligner,VERBE interstitiel,ADJ repos,NOM dépost,null interprétariat,NOM posemètre,NOM intentionner,VERBE indécision,NOM indéfinité,NOM posage,NOM proposer,VERBE interrègne,NOM indéterminable,ADJ interférence,NOM inaction,NOM exposé,ADJ psychologiquement,ADV intermède,NOM possibiliste,NOM ventriculographie,NOM intercession,NOM réentendre,VERBE psychologiser,VERBE reportage,NOM activiste,NOM réexposer,VERBE médium,NOM lignette,NOM réactionnel,ADJ psychogramme,NOM intercalaire,NOM interposer,VERBE dépositaire,NOM approximativement,ADV activiste,ADJ frangé,ADJ définir,VERBE possédé,ADJ biréacteur,NOM psychométrie,NOM interruptible,ADJ apprécier,VERBE déposant,NOM intellectuellement,ADV proportionnement,NOM précieuse,NOM interception,NOM surdéterminer,VERBE positivation,NOM dépouiller,VERBE actionnaire,NOM repoussage,NOM psychiatrie,NOM approché,ADJ médiation,NOM poussade,NOM ininterrompu,ADJ portoire,NOM inintelligible,ADJ médiatement,ADV acte,NOM intentionniste,ADJ intelligentzia,NOM export,NOM rapportable,ADJ importation,NOM tensiorécepteur,NOM possesseur,NOM rompre,VERBE psychanalyse,NOM rapprocher,VERBE positif,ADJ intellectuel,NOM agir,VERBE intermédiat,ADJ dispositif,NOM impost,null déterminisme,NOM portable,ADJ pousser,VERBE tenderie,NOM interpréter,VERBE surintensité,NOM réimposer,VERBE tendresse,NOM apport,NOM intellect,NOM reposement,NOM entendu,ADJ exigeant,ADJ emportement,NOM indéfinissablement,FUNC surexposition,NOM dépose,NOM définissable,ADJ portail,NOM interréaction,NOM dispensation,NOM décompenser,VERBE terminatif,ADJ intercalation,NOM intellectuelle,NOM possible,NOM intention,NOM interfrange,NOM appréciabilité,NOM positif,NOM section,NOM sectionnaire,NOM intervenant,ADJ définition,NOM rompu,ADJ

	intervention,NOM repousseur,NOM positivisme,NOM surexposer,VERBE parapsychologie,NOM inactinique,ADJ exponentielle,NOM interrupteur,ADJ positionnellement,FUNC tensoriel,ADJ intentement,NOM exposition,NOM médiatiser,VERBE hypotenseur,NOM surimposer,VERBE hypertendue,NOM médiumnique,ADJ apporter,VERBE intermédiaire,ADJ intellectualiser,VERBE poussif,ADJ réactance,NOM portuaire,ADJ reportement,NOM reposé,ADJ poser,NOM frangeuse,NOM psychométricienne,NOM intermédiaire,NOM réaction,NOM indisponibilité,NOM intentionnel,ADJ déterminé,ADJ inactifs,NOM portal,ADJ intensité,NOM médioligne,ADJ importance,NOM interférent,ADJ approchant,ADJ suraction,NOM indéterministe,NOM aérotransport,NOM interjeter,VERBE ininterruption,NOM disposer,VERBE portique,NOM intercalaire,ADJ détendu,ADJ active,NOM portant,ADJ transiger,VERBE actinisme,NOM portatif,ADJ intervallaire,ADJ approchable,ADJ portionnaire,NOM rapporté,NOM reposséder,VERBE ventricule,NOM détension,NOM tentateur,NOM préportionné,ADJ prédisposition,NOM soulignement,NOM psychonévrose,NOM superposer,VERBE intellectualiste,NOM apposer,VERBE précieux,ADJ tendancieusement,ADV interfacial,ADJ coactif,ADJ rapprochant,ADJ interrupteur,NOM psychopathologie,NOM déterminatif,ADJ possibilisation,NOM sectionner,VERBE indispensabilité,NOM possibiliste,ADJ possession,NOM poseur,NOM dépouilleur,NOM possessionné,ADJ tendancieux,ADJ indispensable,NOM décisoire,ADJ disponible,ADJ possédant,NOM poser,VERBE décisif,ADJ poussoter,VERBE tenseur,NOM terminage,NOM approximation,NOM réexport,NOM psychagogique,ADJ portage,NOM prédisposer,VERBE possible,ADJ intentionné,ADJ import,NOM prédetermination,NOM indétermination,NOM activateur,NOM intersection,NOM tentation,NOM interfaçage,NOM intermezzo,NOM surinterprétation,NOM psychologisme,NOM inappréciable,ADJ indisponible,ADJ imposable,ADJ rompeur,NOM prédéterminant,NOM psycholinguiste,NOM emport,NOM rapprochement,NOM rapporté,ADJ interjection,NOM intelligibilité,NOM reposant,ADJ réacteur,ADJ tendance,NOM intensive,NOM surcompensation,NOM psychiquement,ADV repoussé,NOM psychonévrosée,NOM intermittence,NOM acticité,NOM médiatisant,ADJ position,NOM intellectualisme,NOM suractivité,NOM tensioactivité,NOM interprétant,NOM positionnel,ADJ positionneur,NOM posséder,VERBE possédable,ADJ psychosociologue,NOM tensioactif,NOM exigibilité,NOM appréciateur,ADJ redéfinir,VERBE intersession,NOM décidé,ADJ repousse,NOM indisposer,VERBE rétropoussette,NOM intellectif,ADJ psychiatriser,VERBE positivité,NOM redéfinition,NOM interventionniste,NOM positiviste,ADJ atermoyer,VERBE dépouillage,NOM psychophysiologie,NOM aligner,VERBE psychagogie,NOM triréacteur,NOM rapproché,NOM détendre,VERBE repoussoir,NOM psychosociologie,NOM dépossession,NOM réexposition,NOM surimposition,NOM actionnement,NOM intempestif,ADJ inintelligiblement,ADV déposséder,VERBE transporté,ADJ détente,NOM inactivité,NOM intentionnalisation,NOM indéterminé,ADJ superstructure,NOM récompenser,VERBE réalignement,NOM portabilité,NOM exaction,NOM appréciablement,ADV tentateur,ADJ tendeur,NOM intervalle,NOM actionner,VERBE porter,VERBE comportement,NOM
29	/1651/ : donnant,ADJ donnable,ADJ dédoublure,NOM dorage,NOM donation,NOM codonataire,ADJ double,NOM surdorer,VERBE bidonne,NOM codonateur,NOM doublure,NOM donne,NOM double,ADJ redoubler,VERBE surdorure,NOM redonder,VERBE redonner,VERBE doublement,NOM dédoublage,NOM doublier,NOM donneuse,NOM doublé,NOM redorer,VERBE donataire,NOM redondance,NOM donneur,NOM doublet,NOM redoublage,NOM doublage,NOM redorure,NOM redondant,ADJ redoublement,NOM redoublante,NOM doubler,VERBE doublon,NOM redorage,NOM donner,VERBE dorure,NOM doubleur,NOM dorer,VERBE doublé,ADJ doubleuse,NOM donateur,NOM dédoublement,NOM doublée,NOM doublement,ADV redoublant,NOM donatrice,NOM doreuse,NOM dédoubler,VERBE doreur,NOM codonataire,NOM don,NOM redoublé,ADJ

30	état
31	/7339/ : querelle,NOM querelleuse,NOM querelleur,NOM querelleur,ADJ quereller,VERBE
32	/1886/ : bruire,VERBE bruissant,ADJ bruiter,VERBE bruit,NOM bruissaillement,NOM bruisser,VERBE bruitage,NOM bruissailler,VERBE bruissement,NOM bruiteur,NOM
33	/4550/ : glorifier,VERBE glorifiable,ADJ glorieux,ADJ glorifiant,ADJ glorieusement,ADV glorificateur,NOM glorieuseté,NOM glorificateur,ADJ glorificatrice,NOM glorification,NOM
34	/309/ : faillibilité,NOM refaçonnement,NOM défaitisme,NOM redéfaire,VERBE défaitiste,ADJ fabrique,NOM affaireux,ADJ fabricatrice,NOM préfabriqué,ADJ méfaire,VERBE refait,NOM fabrication,NOM refaiseuse,NOM défaitiste,NOM factieuse,NOM méfait,NOM faillir,VERBE défaillance,NOM façonnerie,NOM facturer,VERBE fautivement,ADV défaillement,NOM faillite,NOM façonnier,NOM défaçonner,VERBE falloir,VERBE défaillir,VERBE fabricant,NOM affaire,NOM refaiseur,NOM faillie,NOM préfabriqué,NOM affairiste,NOM fabricateur,NOM défait,ADJ fait,NOM faillible,ADJ factionnaire,NOM facturation,NOM préfabrication,NOM refaire,VERBE réfection,NOM refaçonner,VERBE préfabriquer,VERBE fauter,VERBE parfaire,VERBE surfacturer,VERBE refaçonnage,NOM facture,NOM faction,NOM défaillant,ADJ fabriquer,VERBE faute,NOM failli,NOM façonnier,ADJ façon,NOM faire,NOM refabriquer,VERBE factieux,NOM façonnage,NOM faire,VERBE malfaçonné,ADJ défaite,NOM fabricante,NOM façonnière,NOM fabricien,NOM réfectionner,VERBE façonner,VERBE failli,ADJ fautif,ADJ factionnaire,ADJ défaire,VERBE refabrication,NOM défaut,NOM refaçonneur,NOM surfacturation,NOM factieux,ADJ façonnement,NOM
35	lumière
36	/7912/ : scandaleusement,ADV scandaliser,VERBE scandaleux,ADJ scandale,NOM scandalisation,NOM scandalisé,ADJ
37	/1000/ : parlure,NOM parlement,NOM parler,VERBE parlerie,NOM antiparlementaire,NOM reparler,VERBE parler,NOM parlementaire,ADJ parlage,NOM antiparlementaire,ADJ parole,NOM antiparlementarisme,NOM déparler,VERBE préparole,NOM
38	/1856/ : brillantiner,VERBE brillantée,NOM brillement,NOM brillance,NOM brillamment,ADV brillanté,NOM briller,VERBE brillantage,NOM brillant,ADJ brillantine,NOM brillanter,VERBE brillanté,ADJ brillotter,VERBE brilloter,VERBE brillant,NOM
39	/371/ : déraciner,VERBE enraciné,ADJ raciner,VERBE racinienne,NOM racinement,NOM enracinement,NOM déracinage,NOM racinage,NOM enraciner,VERBE racine,NOM déracineur,NOM racinien,NOM indéracinablement,FUNC indéracinable,ADJ raciné,ADJ racinaire,ADJ racinien,ADJ déracinement,NOM
40	/52/ : aboyant,ADJ aboi,NOM aboyer,VERBE aboyeuse,NOM aboyeur,ADJ aboyant,NOM aboyante,NOM aboyeur,NOM aboiement,NOM
41	/6477/ : obtenir,VERBE obtenable,ADJ obtention,NOM
42	/7610/ : tapecu,NOM tape,NOM tapeur,NOM tapette,NOM retapeuse,NOM retaper,VERBE tapager,VERBE retapeur,NOM tapageusement,FUNC retape,NOM tapement,NOM tapeuse,NOM tapecul,NOM retapage,NOM taper,VERBE tapage,NOM tapageur,ADJ
43	/304/ : planièdre,ADJ planitude,NOM planificateur,ADJ plané,ADJ replant,NOM plantaire,ADJ planiste,NOM planté,NOM implantateur,NOM implant,NOM planchéiage,NOM planité,NOM planimètre,NOM implanteur,NOM planeur,NOM

	complanter,VERBE plantain,NOM planant,ADJ planipennes,NOM complantage,NOM plantigrades,NOM planimétrie,NOM plantier,NOM implantable,ADJ monoplan,NOM planteur,NOM plantation,NOM plantigrade,NOM aplani,ADJ plantule,NOM biplan,ADJ planifier,VERBE plante,NOM planoir,NOM planigraphe,NOM planning,NOM planton,NOM planchiste,NOM planéité,NOM planchéier,VERBE planificatrice,NOM déplanter,VERBE planche,NOM déplantoir,NOM complantation,NOM planisme,NOM planimétrique,ADJ plan,ADJ planchette,NOM coplanaire,ADJ plantaginées,NOM aplanissement,NOM implantation,NOM déplantage,NOM plantoir,NOM planification,NOM planter,VERBE implanter,VERBE plantaginacées,NOM planage,NOM planipenne,ADJ plane,NOM aéroplane,NOM planisphère,NOM plançon,NOM plané,NOM plan,NOM planteuse,NOM réimplantation,NOM plant,NOM planificateur,NOM planement,NOM aplanisseuse,NOM triplan,ADJ déplantation,NOM plancher,NOM plantage,NOM aplanisseur,NOM plantaison,NOM aplanir,VERBE diplanthère,NOM planerie,NOM plantigrade,ADJ réimplanter,VERBE planer,VERBE planeur,ADJ planure,NOM complant,NOM planipenne,NOM planigramme,NOM planipennes,ADJ
44	/5375/ : viol,NOM inviolablement,ADV inviolabilité,NOM violer,VERBE violation,NOM violateur,NOM violemment,ADV violenter,VERBE violeuse,NOM violence,NOM violeur,NOM inviolable,ADJ violatrice,NOM violement,NOM violable,ADJ inviolé,ADJ
45	/4051/ : exploser,VERBE explosif,ADJ explosible,ADJ inexplosible,ADJ explosion,NOM

Sémème du mot or

numéro	item
0	/61/ : paralléliseur,NOM antiparasite,NOM dépareillé,ADJ apparaître,VERBE parer,VERBE repassage,NOM parascève,NOM passepied,NOM reparaître,VERBE comparaître,VERBE préparage,NOM pareuse,NOM passerelle,NOM appareil,NOM impassable,ADJ apparition,NOM disparaître,VERBE antiparasite,ADJ passante,NOM apparaux,NOM passavant,NOM repasser,VERBE apparat,NOM repasseur,NOM passéifier,VERBE pareur,NOM appareilleur,NOM passure,NOM paresthésie,NOM dépasser,VERBE apparution,NOM passerine,NOM passager,VERBE parasite,NOM passegrand,ADJ passéisme,NOM passéification,NOM repasseuse,NOM parader,VERBE antiparasitaire,NOM déparer,VERBE dépassant,NOM surpasser,VERBE paraître,VERBE passagère,NOM parage,NOM surpassement,NOM pareil,ADJ insurpassé,ADJ passoire,NOM dépassante,NOM appareiller,VERBE passe,NOM passeur,NOM passation,NOM passéiste,ADJ passade,NOM impréparation,NOM apparent,NOM antiparasitaire,ADJ passegrande,ADJ passement,NOM passériformes,NOM repasse,NOM passager,NOM préparer,VERBE réapparaître,VERBE impréparé,ADJ appareillement,NOM apparoir,VERBE passablement,ADV appareillé,ADJ reparution,NOM passager,ADJ passé,NOM comparoir,VERBE préparation,NOM indépassable,ADJ paradigme,NOM inapparent,ADJ passementer,VERBE parade,NOM passage,NOM dépassement,NOM parution,NOM indépassé,ADJ appareillage,NOM insurpassable,ADJ réapparition,NOM apparaissance,NOM apparence,NOM passée,NOM passant,NOM pas,NOM disparition,NOM passéiste,NOM passette,NOM apparent,ADJ passé,ADJ préparatif,NOM dépassant,ADJ impasse,NOM passereau,NOM passeport,NOM parement,NOM passer,VERBE apparemment,ADV passant,ADJ apparente,NOM comparution,NOM passable,ADJ
1	reconnaître
2	monétaire
3	/7500/ : reflet,NOM reflètement,NOM refléter,VERBE
4	éclat
5	/7424/ : rarement,ADV rare,ADJ
6	/1093/ : rappeleur,NOM rappelé,ADJ appelant,ADJ appeleur,ADJ appellatif,ADJ appellation,NOM appelé,NOM rappeler,VERBE appeler,VERBE rappeleuse,NOM rappelable,ADJ rappel,NOM appel,NOM appeler,NOM appelée,NOM appelante,NOM rappelé,NOM appelé,ADJ appelable,ADJ appelant,NOM appeau,NOM
7	caractère
8	malléable
9	/385/ : feuilletis,NOM feuillet,NOM feuilletonisation,NOM feuilleter,VERBE feuillantine,NOM effeuillaison,NOM feuiller,VERBE feuilletage,NOM feuillette,NOM feuillée,NOM enfeuiller,VERBE feuilleteur,NOM feuilleté,NOM défeuillaison,NOM effeuilleuse,NOM défeuillage,NOM effeuilles,NOM feuillé,NOM effeuilleur,NOM feuillage,NOM feuilleton,NOM feuillagiste,NOM feuillaison,NOM feuilletonniser,VERBE feuilleté,ADJ effeuiller,VERBE feuilleux,ADJ feuillardier,NOM feuillu,ADJ feuille,NOM feuilloler,VERBE effeuillement,NOM défeuillement,NOM défeuiller,VERBE feuillard,NOM effeuillage,NOM préfeuille,NOM feuillé,ADJ feuillade,NOM feuilletement,NOM feuilliste,NOM feuillagé,ADJ feuilletoniser,VERBE feuilletoniste,NOM
10	pépite
11	inaltérable
12	généralement

13	fil
14	natif
15	non
16	/3703/ : enrichissement,NOM richard,NOM richissime,ADJ enrichir,VERBE enrichissant,ADJ richarde,NOM richement,ADV enrichi,ADJ richesse,NOM riche,ADJ
17	épanouissement
18	mou
19	période
20	représenter
21	/2399/ : civilisation,NOM incivilité,NOM civilisable,ADJ civilisé,ADJ civil,NOM civilité,NOM incivil,ADJ civiliser,VERBE incivilisable,ADJ incivilisation,NOM incivilisé,ADJ civilement,ADV civilisateur,ADJ civil,ADJ incivilement,FUNC
22	/679/ : allié,ADJ superalliage,NOM inalliable,ADJ allié,NOM rallye,NOM allier,VERBE interallié,ADJ mésalliance,NOM rallier,VERBE alliée,NOM alliage,NOM alliable,ADJ ralliable,ADJ mésallier,VERBE ralliement,NOM alliance,NOM alliancé,ADJ
23	dureté
24	/3036/ : repenti,ADJ dépeindre,VERBE repeindre,VERBE repeint,NOM empeinturlurer,VERBE peintresse,NOM peinture,NOM peinturlurer,VERBE repenti,NOM peindre,VERBE peint,ADJ peinturlureur,NOM peinturage,NOM peinturlurage,NOM repentie,NOM repentant,NOM repentante,NOM peinturlureuse,NOM peinturer,VERBE peintre,NOM repeinture,NOM peinturier,NOM peinturlure,NOM peintureur,NOM repentir,VERBE repentant,ADJ repentir,NOM peintriot,NOM peinte,ADJ
25	art
26	/126/ : chausson,NOM chaud,NOM chauffer,VERBE réchauffoir,NOM chaudière,NOM préchauffer,VERBE déchaussoir,NOM surchauffe,NOM chaudronnée,NOM chaudronnière,NOM préchauffe,NOM chaudronnier,NOM rechaussement,NOM échangisme,NOM chaussette,NOM échangeur,NOM échangeable,ADJ chausser,VERBE réchauffeur,NOM coéchangiste,NOM chausseterie,NOM réchauffement,NOM déchaussement,NOM chaud,ADJ chaudron,NOM chauffeur,NOM chaudronnerie,NOM chaude,NOM surchauffement,NOM échangiste,ADJ chaussure,NOM chauffoir,NOM réchauffer,VERBE chaufferette,NOM préchauffage,NOM chaudement,ADV réchauffé,ADJ chauffement,NOM échange,NOM inéchangeable,ADJ déchaussé,ADJ chaussonnier,NOM échanger,VERBE rechausser,VERBE surchauffage,NOM réchauffé,NOM réchauffage,NOM chauffé,ADJ rechaussage,NOM déchaussage,NOM échangiste,NOM chauffant,ADJ chaudronné,ADJ chauffage,NOM achaudi,ADJ chaufferie,NOM échangement,NOM surchauffer,VERBE réchaud,NOM chaussant,ADJ chauffe,NOM échangeuse,NOM chaussage,NOM chausseur,NOM déchausser,VERBE chaussetterie,NOM chauffeuse,NOM
27	/3715/ : ensoleillé,ADJ soleillage,NOM solarimètre,NOM soleillée,NOM soleiller,VERBE insolation,NOM solaire,ADJ soleilleux,ADJ soleil,NOM soleillade,NOM insolateur,NOM ensoleillement,NOM solariser,VERBE solarigraphe,NOM insoler,VERBE ensoleiller,VERBE ensoleillage,NOM soleillé,ADJ solarisation,NOM parasoleil,NOM
28	/1430/ : imprévoyance,NOM improvisateur,ADJ entrevoir,VERBE revuiste,NOM improvisade,NOM entrevision,NOM avoir,VERBE imprévue,NOM reviseur,NOM visionneur,NOM impromptu,NOM prévision,NOM improviser,VERBE revisible,ADJ revoir,VERBE avoine,NOM voir,VERBE vue,NOM imprévoyant,ADJ revoyure,NOM ravoir,VERBE prévoyance,NOM improviste,FUNC prévoir,VERBE improviste à l',FUNC

	improvisateur,NOM revisable,ADJ voirie,NOM revision,NOM révision,NOM vision,NOM réviser,VERBE visionnaire,ADJ visionnement,NOM impromptu,ADJ imprévu,NOM prévisible,ADJ visibilité,NOM revue,NOM imprévisibilité,NOM improvisant,ADJ imprévu,ADJ avoiner,VERBE imprévisible,ADJ visible,NOM voir,NOM improvisement,NOM improvisation,NOM avoiné,ADJ imprévision,NOM visionnarisme,NOM visionner,VERBE visionnaire,NOM imprévoyable,ADJ improvisatrice,NOM imprévisiblement,FUNC visionnage,NOM reviser,VERBE
29	luxe
30	/5034/ : imitatif,ADJ imitateur,NOM imitateur,ADJ inimitabilité,NOM inimitable,ADJ inimité,ADJ imitable,ADJ imiter,VERBE imitation,NOM imitabilité,NOM imitatrice,NOM inimitablement,FUNC
31	/5364/ : inutilité,NOM inutiliser,VERBE inutilement,ADV utilitariste,NOM utilité,NOM réutiliser,VERBE utilitairement,FUNC utilisatrice,NOM utilisateur,NOM inutile,ADJ utilitariste,ADJ utilitarisme,NOM utilitaire,ADJ utilisable,ADJ inutilisable,ADJ utile,NOM utilisation,NOM utilisateur,ADJ inutilisé,ADJ utiliser,VERBE utilement,ADV inutilisation,NOM réutilisation,NOM utile,ADJ
32	/4004/ : exceptionnellement,ADV excellentissime,ADJ excepté,ADJ excellence,NOM exception,NOM préexcellence,NOM exceptionnel,ADJ exceptionnel,NOM excepter,VERBE excellent,ADJ excellemment,ADV exceller,VERBE
33	composé
34	/3021/ : nombré,ADJ dénombrement,NOM indénombrable,ADJ nombrant,ADJ dénombrable,ADJ dénombrer,VERBE nombrage,NOM innombrable,ADJ nombreux,ADJ nombreusement,FUNC innombrablement,ADV nombrer,VERBE numératif,ADJ nombrable,ADJ nombre,NOM innombrabilité,NOM numération,NOM
35	/352/ : jauniot,ADJ jaunissure,NOM jaunisse,NOM jaunet,ADJ jaunissement,NOM jaunasse,ADJ jaunissant,ADJ jaune,NOM jaune,ADJ jaunissage,NOM jaunir,VERBE jaunet,NOM jaunâtre,ADJ
36	métal
37	/1067/ : platybasie,NOM plate,NOM platymérie,NOM platée,NOM aplatissement,NOM aplatisseur,NOM platinides,NOM platiner,VERBE platerie,NOM platichlorhydrique,NOM platycéphale,ADJ platine,NOM platoammine,NOM platinose,NOM platymère,ADJ platiné,ADJ aplati,ADJ plat,ADJ platinotypie,NOM platichlorure,NOM aplatissoir,NOM platycnémie,NOM plathelminthes,NOM aplatissant,ADJ platination,NOM platinite,NOM platinoïde,NOM plateau,NOM aplatir,VERBE platiniser,VERBE platinure,NOM platycéphale,NOM aplatissoire,NOM platiammine,NOM platinifère,NOM platitude,NOM platiniridium,NOM platineux,ADJ platière,NOM aplatissage,NOM plat,NOM platinage,NOM aplat,NOM platymère,NOM platycéphalie,NOM platinique,ADJ
38	clarté
39	/2539/ : densimétrique,ADJ densimètre,NOM densifier,VERBE dense,ADJ condensé,ADJ densification,NOM condenser,VERBE densément,ADV densité,NOM surdensité,NOM densimétrie,NOM condensation,NOM
40	/36/ : interposition,NOM interjecter,VERBE abréaction,NOM soulignage,NOM intercalage,NOM intersecter,VERBE port,NOM psychodramatique,ADJ inintelligibilité,NOM lignage,NOM aligneur,NOM transportable,ADJ psychotrope,NOM interruptif,ADJ imposeur,NOM interstice,NOM positionner,VERBE médiatisation,NOM réactiver,VERBE psychométrique,ADJ positivement,ADV psychopathologique,ADJ interfoliage,NOM proposition,NOM terminologie,NOM tendanciellement,FUNC défini,ADJ décidément,ADV terminateur,ADJ psychanalyser,VERBE abrupt,ADJ portière,NOM ventriculostomie,NOM télédétection,NOM inintelligent,ADJ

	transport,NOM pensionnement,NOM indispensable,ADJ entendement,NOM importable,ADJ tenter,VERBE intempestivement,ADV portement,NOM portefeuille,NOM possessif,ADJ possibiliser,VERBE interférométrique,ADJ psychologue,ADJ intelligiblement,ADV aéroport,NOM psycholinguistique,NOM activation,NOM rapport,NOM surpousse,NOM réactivation,NOM intense,ADJ pensionné,NOM abrupt,NOM actif,ADJ abréagir,VERBE exposante,NOM exigible,ADJ lignée,NOM intellectualité,NOM inactivation,NOM rétroactivement,ADV psychiatrique,ADJ médiateur,NOM interruptrice,NOM portatif,NOM souligner,VERBE réactionnaire,ADJ terminaison,NOM déposement,NOM interloquer,VERBE poseur,ADJ psychonévrotique,ADJ reposer,VERBE exporter,VERBE pensionnat,NOM lignager,NOM interpolateur,ADJ indéterminisme,NOM intelligemment,ADV portioncule,NOM inentendu,ADJ porte,NOM psychophysiologiste,NOM hypotenseur,ADJ intensif,NOM enligner,VERBE déposition,NOM repose,NOM approchement,NOM prédétermination,NOM actioniste,NOM sectionnement,NOM tendu,ADJ surcompensé,ADJ psychanalyste,NOM psyché,NOM interligné,ADJ psychopédagogique,ADJ indéfiniment,ADV rétroagissant,ADJ reporteur,NOM intransportable,ADJ surtension,NOM psychothérapie,NOM hypertendu,ADJ effranger,VERBE déterminabilité,NOM repossession,NOM poussée,NOM pension,NOM possibilité,NOM ventriculite,NOM dépouillement,NOM tendue,NOM actinique,ADJ psychographique,ADJ rapporteur,ADJ psychotrope,ADJ possessivité,NOM intermédiate,NOM transporter,VERBE interpolation,NOM atermoyeur,NOM repoussement,NOM psychologiste,ADJ psychisme,NOM aligné,ADJ positionnement,NOM interventionnisme,NOM comporter,VERBE impossibilité,NOM possessoire,NOM tente,NOM indisposé,ADJ surdétermination,NOM emporter,VERBE terminologique,ADJ psycholepsie,NOM pensionner,VERBE importer,VERBE psychogénèse,NOM malintention,NOM imposant,ADJ psychologisation,NOM déterministe,NOM interaction,NOM récompensant,ADJ tentatif,ADJ activeur,NOM alignement,NOM transposer,VERBE indisposition,NOM indéfini,NOM dépotoir,NOM terminateur,NOM appréciateur,NOM mésinterprétation,NOM terminisme,NOM interférométrie,NOM poussement,NOM psycholeptique,ADJ remporter,VERBE porteuse,NOM terminal,NOM suractif,ADJ rétroactes,NOM agissement,NOM indéterminabilité,NOM médiat,ADJ portager,VERBE intermission,NOM approchage,NOM indéterminément,FUNC réagir,VERBE inappréciation,NOM psychosomatique,ADJ exacteur,NOM transigeance,NOM tendu,NOM décider,VERBE ininterprété,ADJ intensif,ADJ tensionnement,NOM exposant,NOM intervenir,VERBE poussette,NOM intellectualisation,NOM alignée,NOM prédéterminer,VERBE hyperactivité,NOM approximatif,NOM entendre,VERBE interpsychologie,NOM indéterministe,ADJ exportatrice,NOM tentement,NOM tentelette,NOM inapprochable,ADJ interpolateur,NOM interventionniste,ADJ intelligence,NOM inapprécié,ADJ compensation,NOM suractivation,NOM actionné,ADJ intello,NOM inappréciablement,FUNC pensionné,ADJ médiumnité,NOM interprétatif,ADJ appréciation,NOM psychanalysé,ADJ psycholeptique,NOM portulan,NOM rapporter,VERBE psychanalysé,NOM interface,NOM inactiver,VERBE indéfinie,NOM repoussant,ADJ intellectuel,ADJ proportionner,VERBE malintentionné,ADJ reposée,NOM psychopédagogue,NOM récompense,NOM médiatisable,ADJ psychothérapeute,NOM porteur,ADJ interminable,ADJ déterminé,NOM approximer,VERBE abruptement,ADV impossible,NOM réactivité,NOM activisme,NOM session,NOM déterministe,ADJ imposition,NOM psychogénique,ADJ tendanciel,ADJ réacteur,NOM distension,NOM interférer,VERBE franger,VERBE appréciatif,ADJ portefeuilliste,NOM apporteur,NOM pensionnaire,NOM portière,ADJ tendage,NOM rompeur,ADJ indéfinitude,NOM intentionnaliser,VERBE psycholinguistique,ADJ portantine,NOM lignerolle,NOM intelligentiel,ADJ portefaix,NOM ligne,NOM positiviste,NOM exposer,VERBE compenser,VERBE actiniquement,ADV porteur,NOM ininterprétable,ADJ tentatrice,NOM possessionnel,ADJ réactivement,FUNC

	intelligentsia,NOM psychologique,ADJ récompenseur,NOM entendeur,NOM intensifier,VERBE porterie,NOM psychogenèse,NOM intenter,VERBE détendeur,NOM intensification,NOM compossibilité,NOM dispenser,VERBE porté,NOM préexponentiel,ADJ psychosomaticien,NOM psychique,ADJ psychopharmacologie,NOM activateur,ADJ interpolatrice,NOM possédante,NOM suractiver,VERBE dépositoire,NOM intensivement,ADV préhypertendu,ADJ psychodiagnostic,NOM imposance,NOM intercéder,VERBE psychiatre,NOM porté,ADJ rapportage,NOM psychotique,ADJ psychanalytique,ADJ positionniste,NOM indéfinissable,ADJ psychophysiologique,ADJ psychosomaticienne,NOM rompeuse,NOM antipsychiatrie,NOM réimposition,NOM poussage,NOM psychosexuel,ADJ prédéterminant,ADJ sectionneur,NOM interprétante,NOM psychométricien,NOM interprétable,ADJ dépens,NOM réactif,NOM actif,NOM lignard,NOM intervenant,NOM portion,NOM interlocutoire,ADJ reporter,NOM appréciable,ADJ psychologie,NOM impossiblement,FUNC activement,ADV proche,ADJ intercepteur,NOM intercepter,VERBE téléreporter,NOM prédéterminisme,NOM mésinterpréter,VERBE tensionner,VERBE supraventriculaire,ADJ rapporteuse,NOM préhypertendue,NOM psychographie,NOM actionnariat,NOM interruption,NOM exportation,NOM terminologue,NOM intelligible,ADJ interminablement,ADV interagir,VERBE interprétation,NOM rompement,NOM interpoler,VERBE psychothérapique,ADJ indéfini,ADJ rupturer,VERBE déposer,VERBE psychotique,NOM exponentiellement,ADV interminé,ADJ interlignage,NOM déterminable,ADJ médiateur,ADJ hypertension,NOM approcher,VERBE exiger,VERBE important,ADJ intempestivité,NOM imposer,VERBE détenteur,NOM frangette,NOM indéterminer,VERBE intermédine,NOM apprécié,ADJ inexigible,ADJ réactrice,NOM précieux,NOM réexporter,VERBE tendreté,NOM terme,NOM repousser,VERBE indécis,ADJ intentionnalité,NOM appréciatrice,NOM intensificateur,NOM déterminant,ADJ poussé,ADJ pensionnée,NOM dépost,NOM frange,NOM intermédiarité,NOM acter,VERBE intentionaliser,VERBE psychoneurologue,NOM disruptif,ADJ report,NOM portemanteau,NOM pousseuse,NOM rétroaction,NOM psychosocial,ADJ métapsychologie,NOM psychologiste,NOM inactif,ADJ interligneur,NOM tende de tranche,NOM reporter,VERBE transigement,NOM prépsychose,NOM inexigibilité,NOM actionnel,ADJ atermoiement,NOM téléreportage,NOM exportateur,NOM tension,NOM rapporteur,NOM activité,NOM exigence,NOM pousseur,NOM transaction,NOM distendre,VERBE inintelligence,NOM possédant,ADJ portier,NOM entestement,null psychomoteur,ADJ psychodrame,NOM intercaler,VERBE reposoir,NOM interjectif,ADJ lignomètre,NOM superposition,NOM intensément,ADV interligne,NOM enlignement,NOM surcompenser,VERBE intellectualisant,ADJ intermittemment,FUNC apposition,NOM positiver,VERBE interlinéaire,ADJ réactif,ADJ psychanalysée,NOM pousse,NOM réexportation,NOM apporteuse,NOM terminé,ADJ ventriculoscopie,NOM déterminant,NOM exportateur,ADJ coaction,NOM intercalement,NOM intentionalité,NOM portelet,NOM psychoprophylaxie,NOM poussoir,NOM frangère,NOM psychasthénie,NOM porte,ADJ hypotension,NOM hypotensif,ADJ psychogène,ADJ polypsychisme,NOM apposement,NOM superposable,ADJ impossible,ADJ décision,NOM interfolier,VERBE intelligent,ADJ interféromètre,NOM psychiatrisée,NOM hypertensif,ADJ entente,NOM interférentiel,ADJ interposé,ADJ psychogérontologue,NOM portoir,NOM interjectionnel,ADJ imposé,ADJ psychogénèse,NOM rétroagir,VERBE transposition,NOM indéterminé,NOM intentionniste,NOM ligner,VERBE psychasthénique,ADJ abréacteur,NOM disposition,NOM psychose,NOM détermination,NOM définisseur,NOM psychopédagogie,NOM ventriculogramme,NOM pose,NOM terminer,VERBE intermédiat,NOM repoussé,ADJ rétrospectivement,ADV activer,VERBE intermittent,ADJ impositionnaire,NOM intercesseur,NOM intentionnellement,ADV poseuse,NOM reposition,NOM tensioactif,ADJ psychobiologie,NOM tendre,VERBE tentative,NOM tensiomètre,NOM ventriculométrie,NOM effrangement,NOM rupture,NOM action,NOM

	psychomotricité,NOM déposante,NOM psychogénie,NOM lignerole,NOM réactimètre,NOM portant,NOM déterminante,NOM hypertendu,NOM psychonévrosé,NOM portionnette,NOM intello,ADJ tensionnage,NOM portée,NOM psychiatrisation,NOM malentendu,NOM réaligner,VERBE transportation,NOM médiatrice,NOM mésentente,NOM interprète,NOM interminis,FUNC interrompre,VERBE approximatif,ADJ psychiatrisé,NOM approche,NOM décompensation,NOM déterminer,VERBE exponentiel,ADJ psychopolynévrite,NOM possiblement,FUNC intellectualiste,ADJ psychologue,NOM rétroactif,ADJ intellection,NOM ventriculaire,ADJ dépôt,NOM psychomoral,ADJ terminal,ADJ transporter,NOM interligner,VERBE interstitiel,ADJ repos,NOM dépost,null interprétariat,NOM posemètre,NOM intentionner,VERBE indécision,NOM indéfinité,NOM posage,NOM proposer,VERBE interrègne,NOM indéterminable,ADJ interférence,NOM inaction,NOM exposé,ADJ psychologiquement,ADV intermède,NOM possibiliste,NOM ventriculographie,NOM intercession,NOM réentendre,VERBE psychologiser,VERBE reportage,NOM activiste,NOM réexposer,VERBE médium,NOM lignette,NOM réactionnel,ADJ psychogramme,NOM intercalaire,NOM interposer,VERBE dépositaire,NOM approximativement,ADV activiste,ADJ frangé,ADJ définir,VERBE possédé,ADJ biréacteur,NOM psychométrie,NOM interruptible,ADJ apprécier,VERBE déposant,NOM intellectuellement,ADV proportionnement,NOM précieuse,NOM interception,NOM surdéterminer,VERBE positivation,NOM dépouiller,VERBE actionnaire,NOM repoussage,NOM psychiatrie,NOM approché,ADJ médiation,NOM poussade,NOM ininterrompu,ADJ portoire,NOM inintelligible,ADJ médiatement,ADV acte,NOM intentionniste,ADJ intelligentzia,NOM export,NOM rapportable,ADJ importation,NOM tensiorécepteur,NOM possesseur,NOM rompre,VERBE psychanalyse,NOM rapprocher,VERBE positif,ADJ intellectuel,NOM agir,VERBE intermédiat,ADJ dispositif,NOM impost,null déterminisme,NOM portable,ADJ pousser,VERBE tenderie,NOM interpréter,VERBE surintensité,NOM réimposer,VERBE tendresse,NOM apport,NOM intellect,NOM reposement,NOM entendu,ADJ exigeant,ADJ emportement,NOM indéfinissablement,FUNC surexposition,NOM dépose,NOM définissable,ADJ portail,NOM interréaction,NOM dispensation,NOM décompenser,VERBE terminatif,ADJ intercalation,NOM intellectuelle,NOM possible,NOM intention,NOM interfrange,NOM appréciabilité,NOM positif,NOM section,NOM sectionnaire,NOM intervenant,ADJ définition,NOM rompu,ADJ intervention,NOM repousseur,NOM positivisme,NOM surexposer,VERBE parapsychologie,NOM inactinique,ADJ exponentielle,NOM interrupteur,ADJ positionnellement,FUNC tensoriel,ADJ intentement,NOM exposition,NOM médiatiser,VERBE hypotenseur,NOM surimposer,VERBE hypertendue,NOM médiumnique,ADJ apporter,VERBE intermédiaire,ADJ intellectualiser,VERBE poussif,ADJ réactance,NOM portuaire,ADJ reportement,NOM reposé,ADJ poser,NOM frangeuse,NOM psychométricienne,NOM intermédiaire,NOM réaction,NOM indisponibilité,NOM intentionnel,ADJ déterminé,ADJ inactifs,NOM portal,ADJ intensité,NOM médioligne,ADJ importance,NOM interférent,ADJ approchant,ADJ suraction,NOM indéterministe,NOM aérotransport,NOM interjeter,VERBE ininterruption,NOM disposer,VERBE portique,NOM intercalaire,ADJ détendu,ADJ active,NOM portant,ADJ transiger,VERBE actinisme,NOM portatif,ADJ intervallaire,ADJ approchable,ADJ portionnaire,NOM rapporté,NOM reposséder,VERBE ventricule,NOM détension,NOM tentateur,NOM préportionné,ADJ prédisposition,NOM soulignement,NOM psychonévrose,NOM superposer,VERBE intellectualiste,NOM apposer,VERBE précieux,ADJ tendancieusement,ADV interfacial,ADJ coactif,ADJ rapprochant,ADJ interrupteur,NOM psychopathologie,NOM déterminatif,ADJ possibilisation,NOM sectionner,VERBE indispensabilité,NOM possibiliste,ADJ possession,NOM poseur,NOM dépouilleur,NOM possessionné,ADJ tendancieux,ADJ indispensable,NOM décisoire,ADJ disponible,ADJ possédant,NOM poser,VERBE décisif,ADJ poussoter,VERBE tenseur,NOM terminage,NOM approximation,NOM

	réexport,NOM psychagogique,ADJ portage,NOM prédisposer,VERBE possible,ADJ intentionné,ADJ import,NOM prédetermination,NOM indétermination,NOM activateur,NOM intersection,NOM tentation,NOM interfaçage,NOM intermezzo,NOM surinterprétation,NOM psychologisme,NOM inappréciable,ADJ indisponible,ADJ imposable,ADJ rompeur,NOM prédéterminant,NOM psycholinguiste,NOM emport,NOM rapprochement,NOM rapporté,ADJ interjection,NOM intelligibilité,NOM reposant,ADJ réacteur,ADJ tendance,NOM intensive,NOM surcompensation,NOM psychiquement,ADV repoussé,NOM psychonévrosée,NOM intermittence,NOM acticité,NOM médiatisant,ADJ position,NOM intellectualisme,NOM suractivité,NOM tensioactivité,NOM interprétant,NOM positionnel,ADJ positionneur,NOM posséder,VERBE possédable,ADJ psychosociologue,NOM tensioactif,NOM exigibilité,NOM appréciateur,ADJ redéfinir,VERBE intersession,NOM décidé,ADJ repousse,NOM indisposer,VERBE rétropoussette,NOM intellectif,ADJ psychiatriser,VERBE positivité,NOM redéfinition,NOM interventionniste,NOM positiviste,ADJ atermoyer,VERBE dépouillage,NOM psychophysiologie,NOM aligner,VERBE psychagogie,NOM triréacteur,NOM rapproché,NOM détendre,VERBE repoussoir,NOM psychosociologie,NOM dépossession,NOM réexposition,NOM surimposition,NOM actionnement,NOM intempestif,ADJ inintelligiblement,ADV déposséder,VERBE transporté,ADJ détente,NOM inactivité,NOM intentionnalisation,NOM indéterminé,ADJ superstructure,NOM récompenser,VERBE réalignement,NOM portabilité,NOM exaction,NOM appréciablement,ADV tentateur,ADJ tendeur,NOM intervalle,NOM actionner,VERBE porter,VERBE comportement,NOM
41	/6056/ : merveillosité,NOM merveilleusement,ADV merveilleux,ADJ
42	/355/ : tisseuse,NOM détisser,VERBE tissulaire,ADJ tissu,NOM tisserande,NOM tissutier,NOM détissage,NOM tissage,NOM tissure,NOM tissé,ADJ tisser,VERBE tissu,ADJ tisseur,NOM tisserin,NOM intissé,NOM tisseranderie,NOM tisserand,NOM
43	conférer
44	/309/ : faillibilité,NOM refaçonnement,NOM défaitisme,NOM redéfaire,VERBE défaitiste,ADJ fabrique,NOM affaireux,ADJ fabricatrice,NOM préfabriqué,ADJ méfaire,VERBE refait,NOM fabrication,NOM refaiseuse,NOM défaitiste,NOM factieuse,NOM méfait,NOM faillir,VERBE défaillance,NOM façonnerie,NOM facturer,VERBE fautivement,ADV défaillement,NOM faillite,NOM façonnier,NOM défaçonner,VERBE falloir,VERBE défaillir,VERBE fabricant,NOM affaire,NOM refaiseur,NOM faillie,NOM préfabriqué,NOM affairiste,NOM fabricateur,NOM défait,ADJ fait,NOM faillible,ADJ factionnaire,NOM facturation,NOM préfabrication,NOM refaire,VERBE réfection,NOM refaçonner,VERBE préfabriquer,VERBE fauter,VERBE parfaire,VERBE surfacturer,VERBE refaçonnage,NOM facture,NOM faction,NOM défaillant,ADJ fabriquer,VERBE faute,NOM failli,NOM façonnier,ADJ façon,NOM faire,NOM refabriquer,VERBE factieux,NOM façonnage,NOM faire,VERBE malfaçonné,ADJ défaite,NOM fabricante,NOM façonnière,NOM fabricien,NOM réfectionner,VERBE façonner,VERBE failli,ADJ fautif,ADJ factionnaire,ADJ défaire,VERBE refabrication,NOM défaut,NOM refaçonneur,NOM surfacturation,NOM factieux,ADJ façonnement,NOM
45	lumière
46	/1276/ : autrement,ADV autrefois,ADV autre,FUNC autre,ADJ
47	décoration
48	/6186/ : monnaie,NOM monnayage,NOM monnayer,VERBE monnayeur,NOM monneron,NOM monnayé,ADJ monnayable,ADJ
49	/2482/ : entretenu,ADJ entreteneuse,NOM entretenage,NOM rétentrice,NOM entretenir,VERBE soutien,NOM soutènement,NOM rétentionnaire,ADJ codétenu,NOM entretènement,NOM rétenteur,ADJ retenir,VERBE détenir,VERBE soutenance,NOM

	codétenue,NOM contenir,VERBE soutenir,VERBE rétention,NOM détention,NOM rétenteur,NOM rétentionniste,NOM entreteneur,NOM retenu,ADJ détenu,NOM contention,NOM tenir,VERBE rétentionnel,ADJ entretien,NOM retenue,NOM rétentionnaire,NOM tenue,NOM
50	or
51	/117/ : certificateur,NOM acertainer,VERBE certifié,ADJ certainement,ADV certifier,VERBE certain,ADJ certificat,NOM certification,NOM
52	/7062/ : plus ou moins,ADV plus de,ADV plus,ADV plusieurs,FUNC
53	étalon
54	/2728/ : cuivrerie,NOM cuivré,ADJ cuivrique,ADJ cuivrer,VERBE cuivreux,ADJ cuivre,NOM cuivrage,NOM
55	majorité
56	précieux
57	émail
58	/3833/ : soie,NOM soierie,NOM
59	/7043/ : pleine,ADJ plein,ADJ pleine,FUNC pleine,NOM
60	/401/ : proportion,NOM disproportionnement,NOM propriétaire,NOM disproportion,NOM propreté,NOM disproportionner,VERBE expropriatrice,NOM appropriable,ADJ appropriatif,ADJ appropriage,NOM propre,ADJ apprendre,VERBE appris,NOM expropriation,NOM exproprier,VERBE réapprendre,VERBE apprentie,NOM copropriétaire,NOM copropriété,NOM malpropreté,NOM approprié,ADJ réapprentissage,NOM propriété,NOM disproportionné,ADJ expropriateur,NOM apprentissage,NOM proportionné,ADJ propre,NOM approprier,VERBE apprenti,NOM apprise,NOM malappris,NOM appropriement,NOM appropriation,NOM appris,ADJ malpropre,ADJ expropriateur,ADJ
61	/1101/ : preneur,NOM entr'ouvrir,VERBE reproductibilité,NOM décomposant,ADJ reprocheur,ADJ productif,ADJ composant,ADJ représenté,ADJ produire,VERBE improduit,ADJ incompréhensiblement,ADV entrouvrir,VERBE reproductivité,NOM mécomprendre,VERBE production,NOM rentré,NOM entrance,NOM représentation,NOM surprise,NOM appréhendement,NOM emprisonné,ADJ autoreproducteur,ADJ rentrayeur,NOM rentrant,NOM reprisage,NOM mécompréhension,NOM reproductif,ADJ incompréhensible,ADJ apprésentation,NOM prisonnier,NOM compréhension,NOM entr'ouvrement,NOM reproduire,VERBE plexus,NOM reproductrice,NOM pris,ADJ sentimentaliste,NOM sentimentalité,NOM surproduction,NOM entrouverture,NOM entreprise,NOM indécomposé,ADJ déprise,NOM reprographique,ADJ senti,NOM incomplexe,ADJ sentimentalisation,NOM irreprésentable,ADJ rentrante,NOM incompréhensif,ADJ reproche,NOM prisonnière,NOM rentrayeuse,NOM présent,ADJ prison,NOM représentée,NOM prise,NOM représenter,VERBE présenter,VERBE repriser,VERBE incompris,ADJ décomposer,VERBE reprise,NOM reprocher,VERBE présence,NOM procès,NOM compréhensible,ADJ représentante,NOM preneuse,NOM reproductivement,ADV entrer,VERBE reprisable,ADJ rentrant,ADJ représentativité,NOM reprographier,VERBE prendre,VERBE rentrage,NOM indécomposable,ADJ coproduction,NOM déprendre,VERBE repriseur,ADJ reprochable,ADJ reprendre,VERBE imprenable,ADJ composante,NOM présentation,NOM reproductible,ADJ improductivement,FUNC repriseuse,NOM coproduire,VERBE improductif,ADJ preneur,ADJ producteur,NOM répréhension,NOM préhension,NOM représentable,ADJ senti,ADJ entrant,NOM comprendre,VERBE reproduction,NOM rentrure,NOM entreprendre,VERBE incompréhensibilité,NOM appréhension,NOM sentimentaliser,VERBE produit,ADJ surproduit,NOM improductivité,NOM sentiment,NOM complexe,ADJ représenté,NOM

	représentatif,ADJ sentimental,ADJ prisonnier,ADJ produit,NOM reproducteur,ADJ intercompréhension,NOM sentimentalisme,NOM entrée,NOM rentrée,NOM sentir,VERBE emprisonner,VERBE dissentiment,NOM surproduire,VERBE appréhender,VERBE rentrer,VERBE rentré,ADJ surreprésentation,NOM apprésenter,VERBE entrant,ADJ présent,NOM composant,NOM représentativement,FUNC reproducteur,NOM incompréhension,NOM décomposition,NOM entrante,NOM entrepreneur,NOM emprisonnement,NOM surprendre,VERBE appréhendé,ADJ représentant,NOM reprographie,NOM décomposable,ADJ
62	/2909/ : galon,NOM galonné,ADJ dégalonner,VERBE galonnage,NOM galonner,VERBE
63	très
64	évoquer
65	eau
66	/1856/ : brillantiner,VERBE brillantée,NOM brillement,NOM brillance,NOM brillamment,ADV brillanté,NOM briller,VERBE brillantage,NOM brillant,ADJ brillantine,NOM brillanter,VERBE brillanté,ADJ brillotter,VERBE brilloter,VERBE brillant,NOM
67	/4275/ : infortuné,ADJ fortuité,NOM infortune,NOM fortuitement,ADV fortuné,ADJ fortuit,ADJ fortune,NOM
68	air
69	/967/ : atomiser,VERBE atomiste,ADJ atomiste,NOM atomique,ADJ polyatomique,ADJ atomisé,ADJ atomisation,NOM interatomique,ADJ atomistique,ADJ diatomique,ADJ triatomique,ADJ antiatomique,ADJ atomiquement,ADV atome,NOM atomiseur,NOM atomicité,NOM monoatomique,ADJ subatomique,ADJ atomisme,NOM atomistique,NOM atomicien,NOM
70	préparation
71	/208/ : décolletage,NOM couvert,ADJ décoratrice,NOM découverte,NOM décolleté,NOM décorateur,NOM décoration,NOM recouverture,NOM découvert,ADJ découvreur,NOM couvrement,NOM découvrir,VERBE couvrant,ADJ recouvrer,VERBE décorativement,ADV redécouverte,NOM décoré,ADJ adécoratif,ADJ couvrure,NOM décolleteur,NOM recouvrir,VERBE décolleteuse,NOM recouvrable,ADJ décolleté,ADJ décorer,VERBE couverture,NOM recouvrement,NOM découverture,NOM collerette,NOM couvert,NOM décolleter,VERBE couverte,NOM couvrir,VERBE découvert,NOM décor,NOM couvrante,NOM collet,NOM recouvrance,NOM couvercle,NOM décorés,NOM décorum,NOM recouvrage,NOM redécouvrir,VERBE découvrement,NOM couvraine,NOM décoratif,ADJ
72	présenter
73	/4149/ : nickeline,NOM nickeler,VERBE nickelage,NOM nickelifère,ADJ nickelure,NOM nickel,NOM nickelé,ADJ
74	/533/ : transformer,VERBE formaliserse,VERBE fondant,ADJ profonde,NOM formaliste,ADJ profond,NOM profondeur,NOM néoformation,NOM formolage,NOM formuler,VERBE refondage,NOM préforme,NOM cofondatrice,NOM fortiori,FUNC forte,ADV fonder,VERBE biforme,ADJ reforming,NOM réformette,NOM préformant,ADJ fusionisme,NOM fondamentaliste,ADJ déformé,ADJ formalisé,ADJ préformé,ADJ formaliser,VERBE fusionner,VERBE forme,NOM forte,ADJ formulaire,NOM efforcement,NOM informatique,ADJ fondage,NOM formateur,NOM conformer,VERBE fuser,VERBE formulique,ADJ fondre,VERBE déformable,ADJ formulation,NOM préformage,NOM fondement,NOM fonderie,NOM informatrice,NOM format,NOM déformer,VERBE informaticien,NOM formateur,ADJ fondateur,NOM

	réformer,VERBE approfondi,ADJ approfondisseur,NOM fondée,NOM fondamentalité,NOM réformisme,NOM fusionnement,NOM réformiste,NOM déformation,NOM formeur,NOM formier,NOM formolateur,NOM surinformation,NOM périinformatique,NOM déformateur,ADJ fondu,NOM fusionniste,ADJ informatisation,NOM réformiste,ADJ informé,NOM transformation,NOM réformé,NOM cofondateur,NOM conforme,ADJ formellement,ADV informité,NOM formolisation,NOM formol,NOM formiate,NOM informatif,ADJ informaticienne,NOM téléinformatique,NOM déformant,ADJ informationnel,ADJ confusionnisme,NOM effondrer,VERBE parfondre,VERBE information,NOM préformation,NOM réformé,ADJ réformée,NOM informel,ADJ informé,ADJ fusionnage,NOM approfondir,VERBE refonte,NOM informatiser,VERBE infondé,ADJ informant,ADJ refusion,NOM confondre,VERBE formant,NOM refondre,VERBE info,NOM fondé,ADJ réformation,NOM réformateur,NOM formage,NOM approfondissant,ADJ formatrice,NOM méforme,NOM informateur,NOM formant,ADJ réformatrice,NOM formolé,ADJ effondrement,NOM formique,ADJ formalisable,ADJ reformage,NOM formaliste,NOM approfondissement,NOM réforme,NOM informer,VERBE conformateur,NOM formulable,ADJ fortiori a,FUNC fondamentaliste,NOM reformation,NOM fondue,NOM fusion,NOM reformulation,NOM confusion,NOM fondatrice,NOM informulable,ADJ formalité,NOM fondamental,ADJ fondé,NOM formoler,VERBE effondrilles,NOM efforcer,VERBE fondoir,NOM profond,ADJ irréformable,ADJ effort,NOM formalisant,ADJ fortement,ADV informatisé,ADJ fort,ADJ formalisme,NOM préformer,VERBE irréformabilité,NOM fondamentalement,ADV reformuler,VERBE reformer,VERBE formel,ADJ superforme,NOM fonte,NOM informulé,ADJ fusage,NOM infondre,VERBE uniformément,ADV former,VERBE réformateur,ADJ conformation,NOM indéformabilité,NOM fondeur,NOM informatique,NOM fond,NOM formatif,ADJ réformage,NOM fondant,NOM formation,NOM fondation,NOM formalisation,NOM fondu,ADJ fusionnisme,NOM indéformable,ADJ formule,NOM profondément,ADV informateur,ADJ conformément,ADV informe,ADJ
75	considérer
76	/6458/ : objectité,NOM objectivation,NOM objectivable,ADJ objet,NOM objectif,NOM objectiver,VERBE objectif,ADJ objectal,ADJ téléobjectif,NOM
77	/2621/ : coule,NOM couleur,NOM
78	/2329/ : chosette,NOM chosification,NOM chosisme,NOM choséifier,VERBE chose,NOM chosifier,VERBE chosmer,VERBE
79	ductile
80	/432/ : grandirostre,ADJ grandissime,ADJ grande,ADJ grandette,ADJ grandeur,NOM grandissant,ADJ agrandissement,NOM grand,ADJ grandir,VERBE agrandissant,ADJ grandissement,NOM grandelet,ADJ grands,NOM grandiose,ADJ grandesse,NOM grandi,ADJ supergrand,NOM agrandir,VERBE agrandisseur,NOM grandiosement,FUNC grandirostre,NOM grandement,ADV
81	/2644/ : invariance,NOM variété,NOM variolé,ADJ variétal,ADJ covariant,ADJ covariation,NOM variable,ADJ covariante,NOM invariable,ADJ variance,NOM variomètre,NOM variolique,ADJ varioleux,ADJ varier,VERBE variationnel,ADJ variabilité,NOM invariablement,ADV varioleuse,NOM varioloïde,NOM variable,NOM invariant,ADJ varié,ADJ invariant,NOM variole,NOM covariant,NOM variolisation,NOM varia,NOM varioleux,NOM variation,NOM invariabilité,NOM variocoupleur,NOM covariance,NOM monovariant,ADJ
82	/3030/ : rempailleur,NOM empailleuse,NOM empaillage,NOM rempaillage,NOM empaillé,ADJ paillé,NOM pailletage,NOM paille,NOM dépaillage,NOM empailler,VERBE paillet,NOM paillassonner,VERBE rempailler,VERBE paillade,NOM paillassine,NOM pailler,NOM empaillée,NOM pailletement,NOM pailleux,ADJ

	paillotte,NOM pailleteur,NOM empaillement,NOM paillère,NOM pailleter,VERBE pailleur,NOM paillasse,NOM paillote,NOM empaillé,NOM paillassonnage,NOM paillis,NOM pailleuse,NOM rempailleuse,NOM paillot,NOM pailleté,ADJ paillat,NOM empailleur,NOM pailleté,NOM paillon,NOM paillage,NOM paillasson,NOM pailler,VERBE dépailler,VERBE paillé,ADJ paillasserie,NOM paillette,NOM
83	/803/ : argentométrie,NOM argenteur,NOM argentable,ADJ argentomètre,NOM argentine,NOM argenteuse,NOM argentan,NOM argentopyrite,NOM argentiste,NOM argentobismuthite,NOM argenter,VERBE argenté,ADJ argentage,NOM argentique,ADJ argentier,NOM argentophile,ADJ argentement,NOM argenton,NOM argent,NOM argenteux,ADJ argentojarosite,NOM argentin,ADJ argenterie,NOM argenture,NOM argentifère,ADJ argentation,NOM
84	/321/ : filatrice,NOM fileuse,NOM défilage,NOM affiler,VERBE filature,NOM filateur,NOM filer,VERBE filage,NOM défiler,VERBE affiloire,NOM fileté,ADJ défilé,NOM filement,NOM fileter,VERBE affile,NOM filetage,NOM filet,NOM affileuse,NOM parfiler,VERBE défilé,ADJ affiloir,NOM fileur,NOM parfilage,NOM monofilament,NOM file,NOM affilage,NOM défilement,NOM affileur,NOM affile d',FUNC entrefilet,NOM filerie,NOM filaturer,VERBE défilade,NOM affilé,ADJ
85	/3264/ : solutionner,VERBE résorbable,ADJ résorber,VERBE solucamphre,NOM résoudre,VERBE solutionnaire,NOM insoluble,ADJ résolvant,ADJ solutionnement,NOM résolue,NOM irrésolution,NOM solubiliser,VERBE résolvante,NOM dissoudre,VERBE solutionniste,NOM solutionniste,ADJ insolubilisation,NOM indissolubilité,NOM insolubilité,NOM dissolution,NOM résolu,NOM soluté,NOM résolutif,ADJ résolu,ADJ résorption,NOM résolutoire,ADJ solubilisation,NOM indissoluble,ADJ résoluble,ADJ résolubilité,NOM résolution,NOM solution,NOM insolubiliser,VERBE irrésolu,ADJ soluble,ADJ solubilité,NOM
86	état
87	/5788/ : lune,NOM luneux,ADJ
88	/138/ : acidosique,ADJ acidimétrie,NOM acidimétrique,ADJ acidimètre,NOM aciduler,VERBE acide,ADJ acidifier,VERBE acidification,NOM acidage,NOM acidifiant,ADJ acidose,NOM acide,NOM acidifère,ADJ acidifiable,ADJ acidulation,NOM acidule,ADJ acidité,NOM monoacide,ADJ acidogène,ADJ peracide,NOM acidulé,ADJ acidifié,ADJ acidoïde,ADJ
89	/1870/ : brodeur,NOM brodage,NOM surbroder,VERBE broder,VERBE rebroder,VERBE broderie,NOM brodé,ADJ brodequin,NOM surbrodage,NOM brodeuse,NOM
90	/854/ : surpuissance,NOM puissance,NOM puissamment,ADV puissant,ADJ

A3) Cotextes du corpus de contes

1er cotexte : nacre (1289 familles de traits sémantiques)

Non loin de Smyrne, sous les hauts platanes, là où le marchand pousse ses chameaux chargés de marchandises qui lèvent fièrement leurs longs cous et foulent maladroitement la terre sacrée, j'ai vu une haie de rosiers en fleurs. Des pigeons sauvages volaient entre les branches des hauts arbres et leurs ailes scintillaient dans les rayons de soleil comme si elles étaient **nacrées**.

2e cotexte : nacre et sable (1329 familles de traits sémantiques)

C'était un petit lac limpide qui ressemblait à un diamant vert enchâssé dans un anneau de fleurs, et où se jouaient des poissons de toutes les nuances de l'orange et de la cornaline, des carpes de Chine couleur d'ambre, des cygnes blancs et noirs, des sarcelles exotiques vêtues de pierreries, et, au fond de l'eau, des coquillages de **nacre** et de pourpre, des salamandres aux vives couleurs et aux panaches dentelés, enfin tout un monde de merveilles vivantes glissant et plongeant sur un lit de **sable** argenté, où poussaient des herbes fines, plus fleuries et plus jolies les unes que les autres.

3e cotexte : sable (1119 familles de traits sémantiques)

- Un plat très nécessaire à ta pauvre petite existence, répondit-elle ; je fais du granit, c'est-à-dire qu'avec la poussière je fais la plus dure et la plus résistante des pierres. Il faut bien cela, pour enfermer le Cocyte et le Phlégéthon. Je fais aussi des mélanges variés des mêmes éléments. Voici ce qu'on t'a montré sous des noms barbares, les gneiss, les quartzites, les talcschistes, les micaschistes, etc. De tout cela, qui provient de mes poussières, je ferai plus tard d'autres poussières avec des éléments nouveaux, et ce seront alors des ardoises, des **sables** et des grès. Je suis habile et patiente, je pulvérise sans cesse pour réagglomérer. La base de tout gâteau n'est-elle pas la farine ? Quant à présent, j'emprisonne mes fourneaux en leur ménageant toutefois quelques soupiraux nécessaires pour qu'ils ne fassent pas tout éclater. Nous irons voir plus haut ce qui se passe. Si tu es fatiguée, tu peux faire un somme, car il me faut un peu de temps pour cet ouvrage.

4e cotexte : sable (510 familles de traits sémantiques)

Après avoir marché assez longtemps sur le **sable**, il se baissa et écrivit ces vers avec une canne qu'il tenait dans sa main :

5e cotexte : pollen (559 familles de traits sémantiques)

De la maison du gouverneur indigène, où la mère et le faux enfant avaient couché, un bourjane apporta triomphalement dans une soubika les déjections du petit. Raketaka avait fabriqué des ordures d’enfant avec de la patate cuite colorée par du **pollen** de citrouille. Sauf l’odeur, c’était à s’y méprendre.

6e cotexte : rose (739 familles de traits sémantiques)

Des buissons de **roses** de toutes nuances et de tous parfums se miraient dans l'eau, ainsi que le fût des colonnes et les belles statues de marbre de Paros placées sous les arcades. Au milieu du bassin jaillissait en mille fusées de diamants et de perles un jet d'eau qui retombait dans de colossales vasques de nacre.

7e cotexte : rose (1123 familles de traits sémantiques)

Une échelle, dont je ne pouvais apercevoir ni la base ni le faîte, se présentait en effet devant nous. Je suivis la fée et me trouvai avec elle dans les ténèbres, mais je m'aperçus alors qu'elle était toute lumineuse et rayonnait comme un flambeau. Je vis donc des dépôts énormes d'une pâte rosée, des

blocs d'un cristal blanchâtre et des lames immenses d'une matière vitreuse noire et brillante que la fée se mit à écraser sous ses doigts ; puis elle pila le cristal en petits morceaux et mêla le tout avec la pâte **rose**, qu'elle porta sur ce qu'il lui plaisait d'appeler un feu doux.

8e cotexte : rose (500 familles de traits sémantiques)

Dans tous les chants d'Orient on parle de l'amour du rossignol pour la **rose**. Dans les nuits silencieuses, le troubadour ailé chante sa sérénade à la fleur suave.

9e cotexte : rose (568 familles de traits sémantiques)

Il cueillit la **rose**, l'inséra dans son livre et l'emporta ainsi sur un autre continent, dans son pays lointain. La **rose** fana de chagrin et demeura aplatie dans le livre. Lorsque le chanteur revint chez lui, il ouvrit le livre et dit : Voici une **rose** de la tombe d'Homère.

10e cotexte : éclat et or (660 familles de traits sémantiques)

Là-dessus, elle s'éloigna en poussant un grand **éclat** de rire, et il me sembla la voir se dissoudre et s'élever en grande traînée d'**or**, rougi par le soleil couchant

11e cotexte : éclat (435 familles de traits sémantiques)

Des bruits formidables, des sifflements aigus, des explosions, des **éclats** de tonnerre remplissaient cette caverne de nuages noirs où je me sentais enfermée.

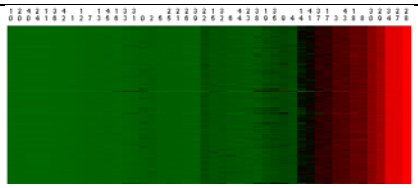
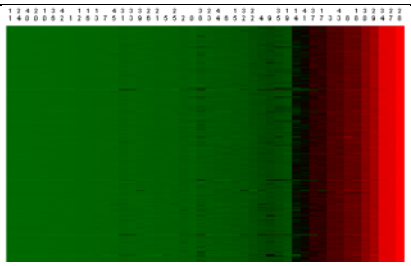
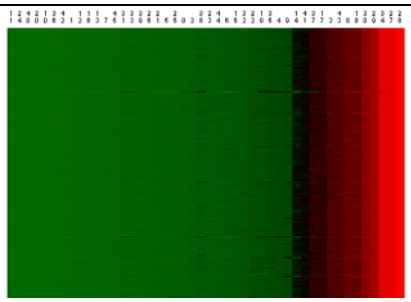
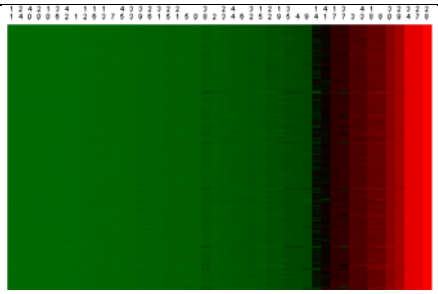
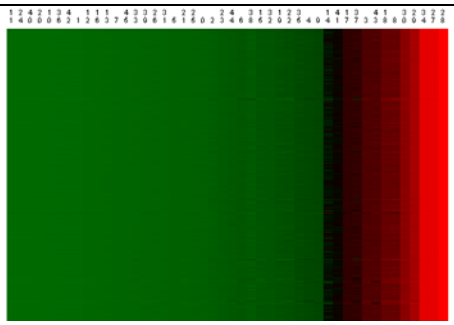
12e cotexte : fer (602 familles de traits sémantiques)

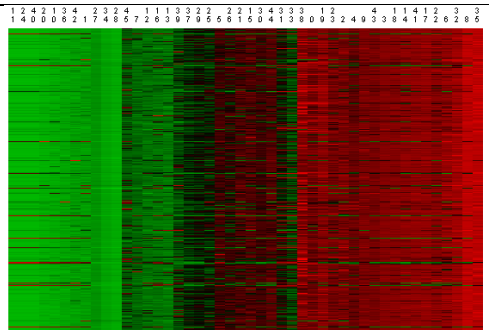
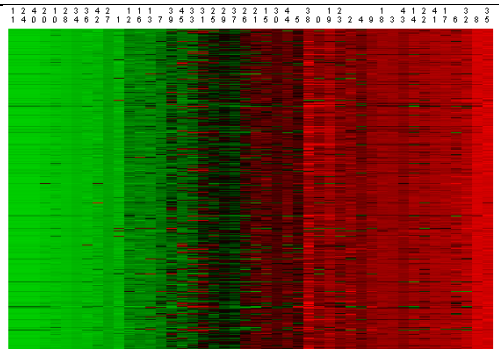
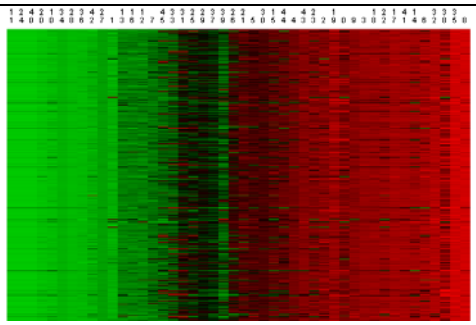
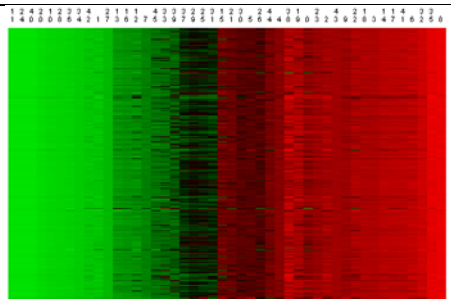
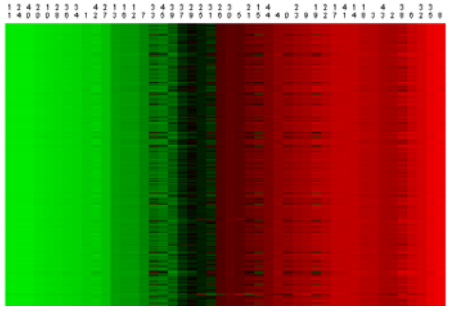
La princesse poussa un cri terrible en apercevant le nain mais ses plaintes ne servirent qu'à aigrir ce petit monstre : avec deux mots de son grimoire, il fit paraître deux géants qui chargèrent le roi de chaînes et de **fers**.

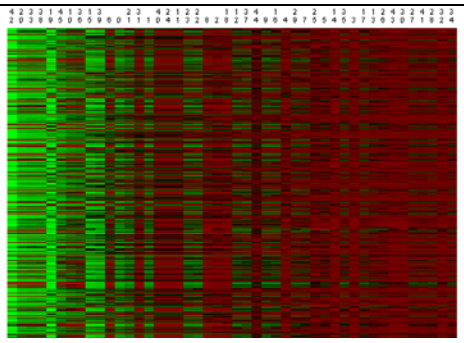
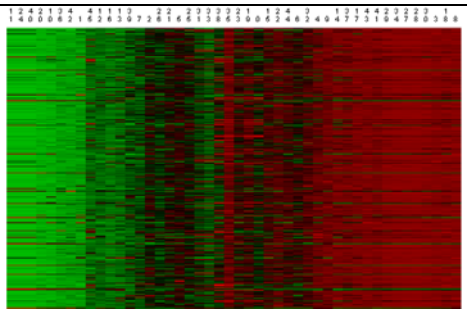
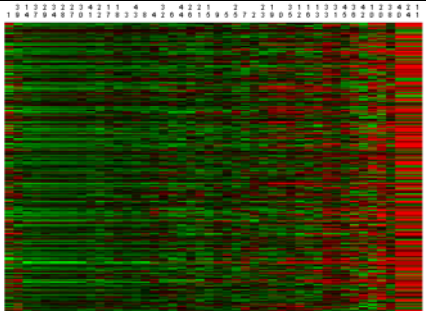
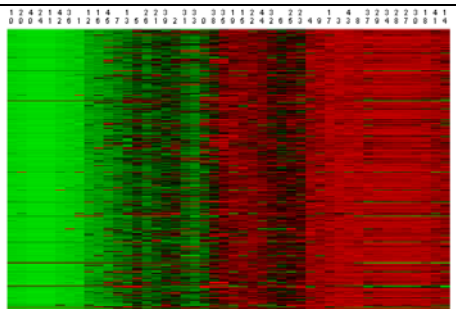
13e cotexte : fer (1654 familles de traits sémantiques)

Le lendemain matin, Margot sortit remplir le seau, le suspendit dans la cheminée et alluma le feu. "Nous allons d'abord faire du pain" dit la vieille, "j'ai déjà chauffé le four et pétri la pâte." Elle poussa la pauvre Margot vers le four duquel les flammes déjà sortaient. "Penche toi et vois si c'est suffisament chaud afin que nous puissions y enfourner le pain." Puis lorsque Margot fut assez proche, elle voulut ouvrir le four pour la faire rôtir dedans et ensuite la dévorer. Mais Margot devinant ses intentions dit : "Je ne sais pas comment faire pour entrer dedans !" "Oie stupide," dit la vieille, "la porte est assez grande, ne vois-tu pas que même moi je peux y passer" affirma-t-elle en rampant et en passant la tête dans le four. Alors Margot lui donna un bon coup si bien qu'elle bascula dedans puis elle referma la porte en **fer** et tira le verrou. "Hou ! hou !" hurla-t-elle horriblement ; Margot partit en courant tandis que l'horrible sorcière brûlait abominablement.

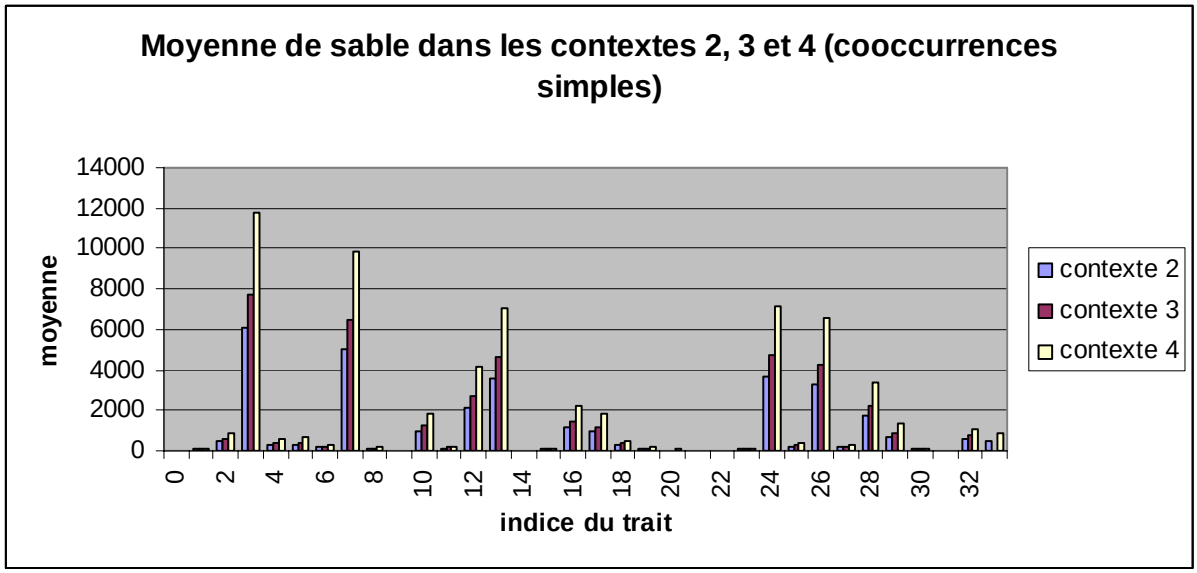
A4) Comparaison de transformations mathématiques : exemple d'éclat dans le contexte n°10

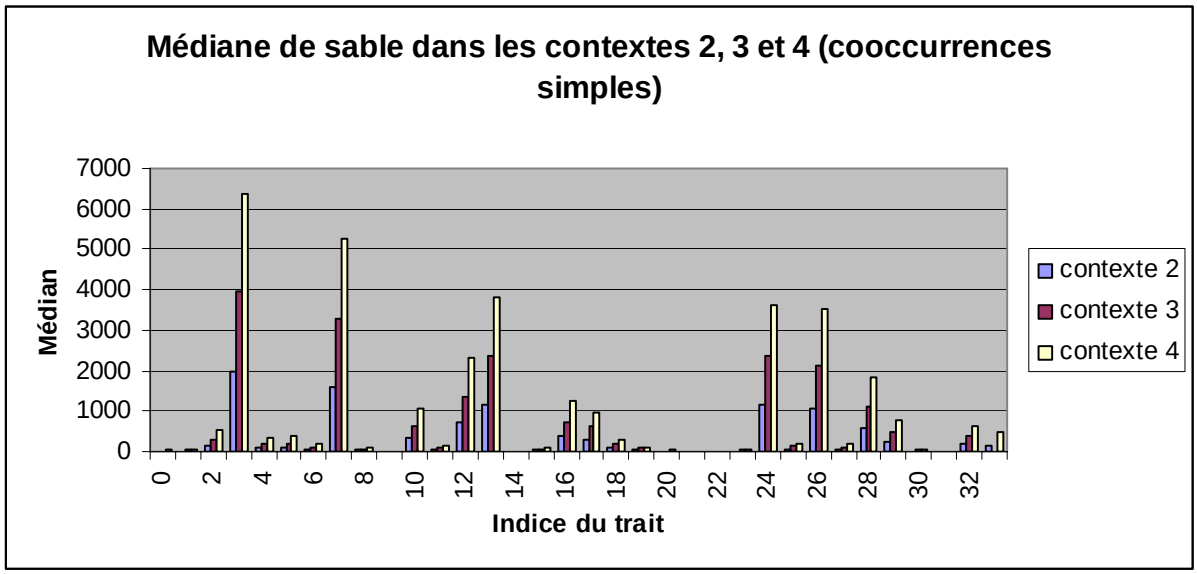
Transformation	Matrice
matrice des cooccurrences sans transformation	
méthode adaptée de LSA, projection sur 50 dimensions	
méthode adaptée de LSA, projection sur 25 dimensions	
méthode adaptée de LSA projection sur 10 dimensions	
méthode adaptée de LSA, projection sur 5 dimensions	

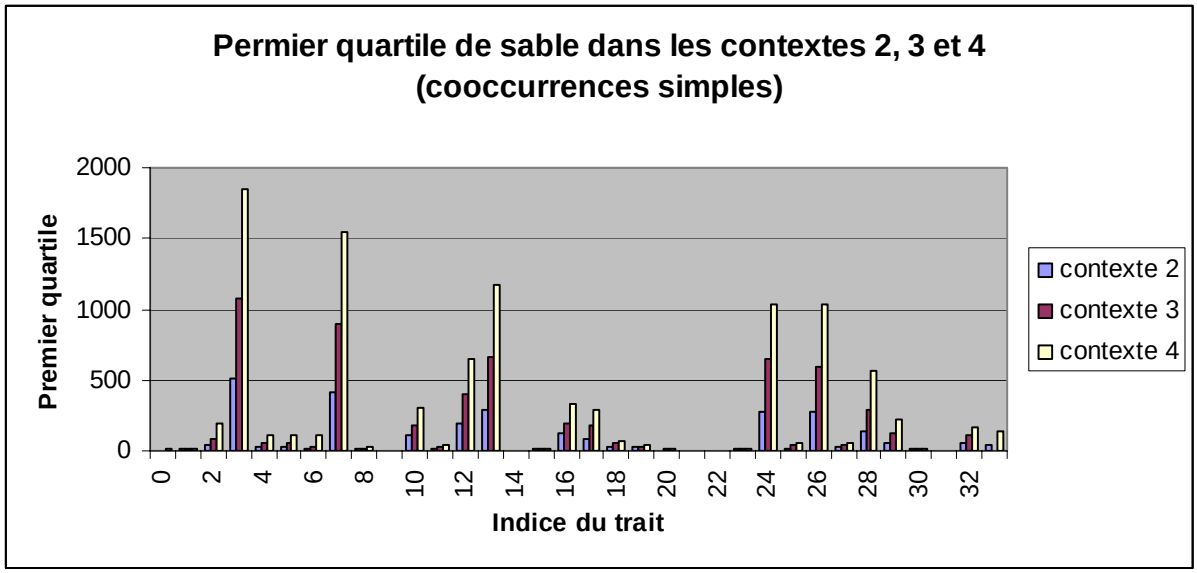
tf-idf	
tf-idf puis méthode adaptée de LSA, projection sur 50 dimensions	
tf-idf puis méthode adaptée de LSA, projection sur 25 dimensions	
tf-idf puis méthode adaptée de LSA, projection sur 10 dimensions	
tf-idf puis méthode adaptée de LSA, projection sur 5 dimensions	

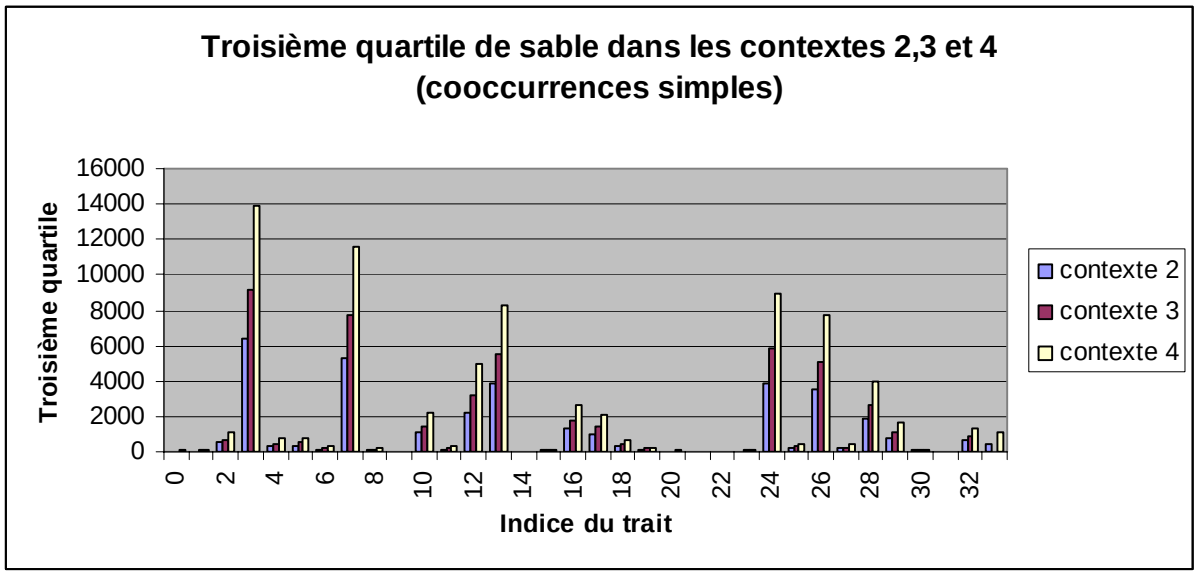
tf-idf puis méthode adaptée de LSA, projection sur 50 dimensions, puis calcul de la matrice des cosinus (pas de produit de la matrice d'occurrences par sa transposée)	
matrice des cosinus	
Calcul des cooccurrences par produit de la matrice d'occurrences par sa transposée, puis application de la méthode adaptée du χ^2	
Application de la méthode adaptée du χ^2 puis calcul des cooccurrences par produit de la matrice d'occurrences par sa transposée	

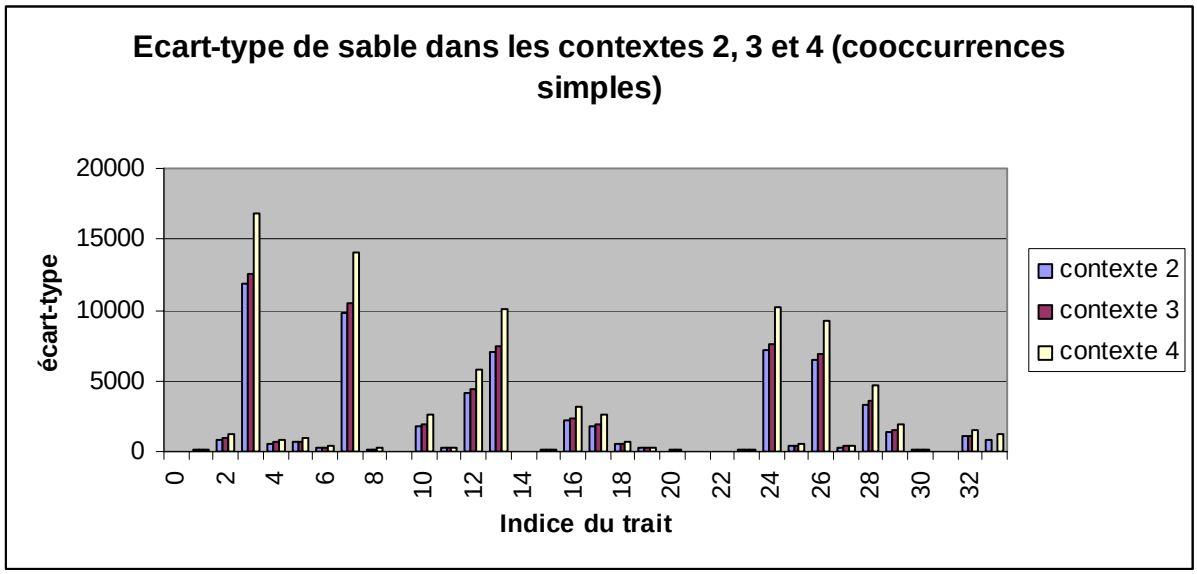
A5) Comparaison de contextes : indicateurs de valeurs centrales et de dispersion du mot sable

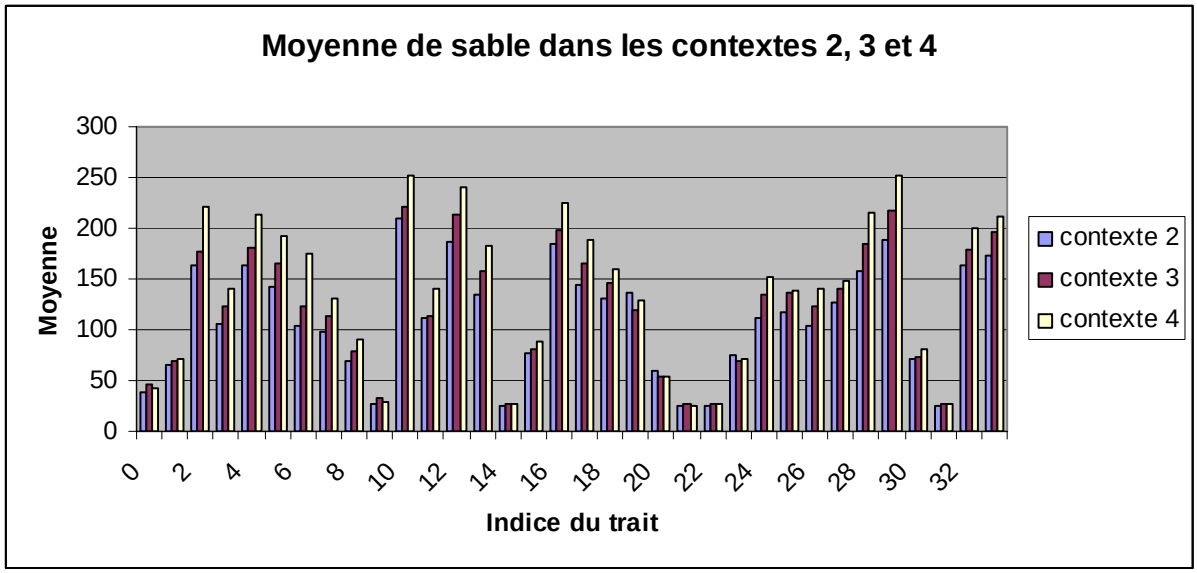


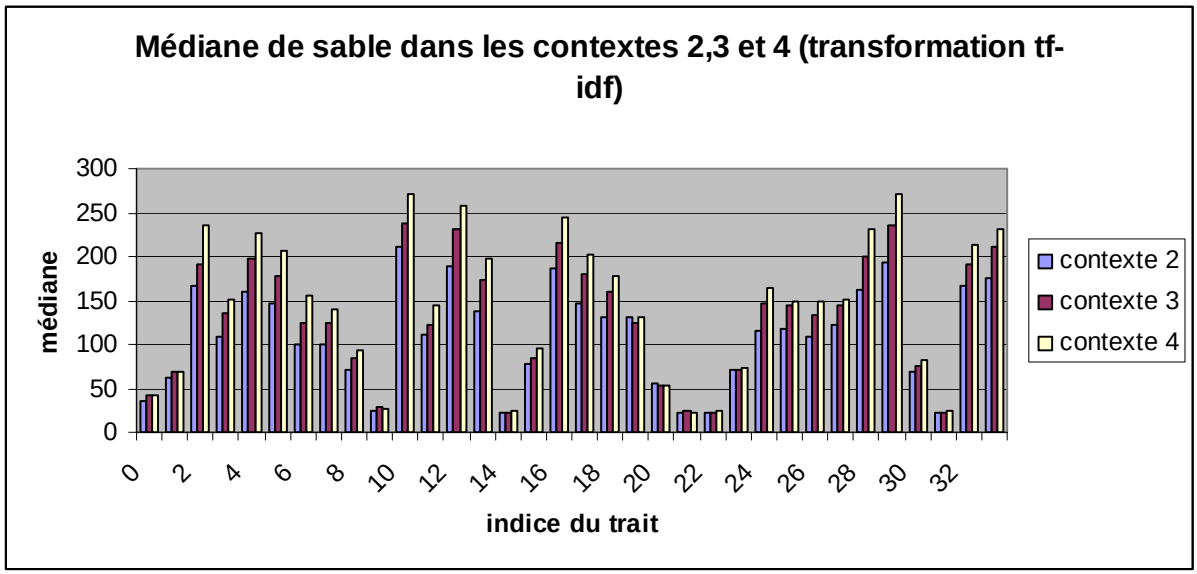


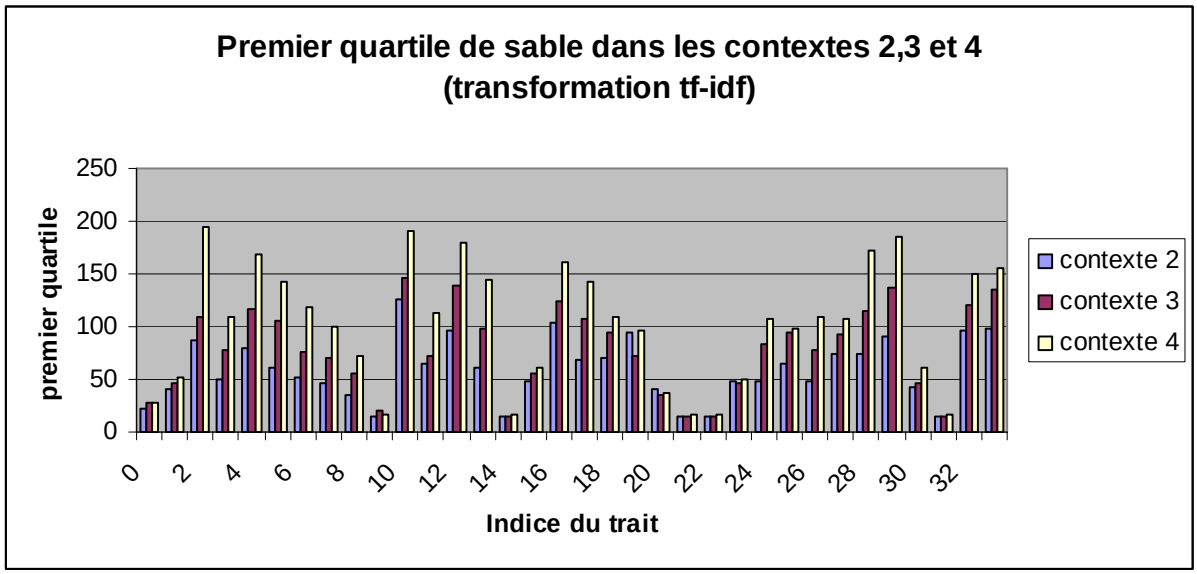

Permier quartile de sable dans les contextes 2, 3 et 4 (cooccurrences simples)
Premier quartile
2000
1500
1000
500
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32
Indice du trait
contexte 2
contexte 3
contexte 4

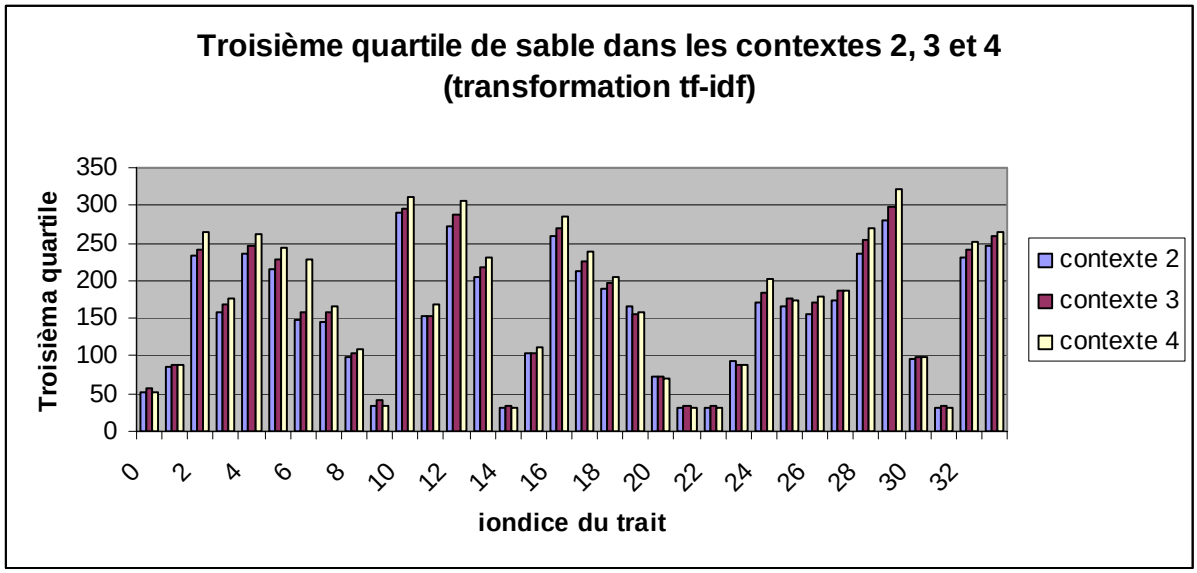

Troisième quartile de sable dans les contextes 2,3 et 4 (cooccurrences simples)
Troisième quartile
16000
14000
12000
10000
8000
6000
4000
2000
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32
Indice du trait
contexte 2
contexte 3
contexte 4

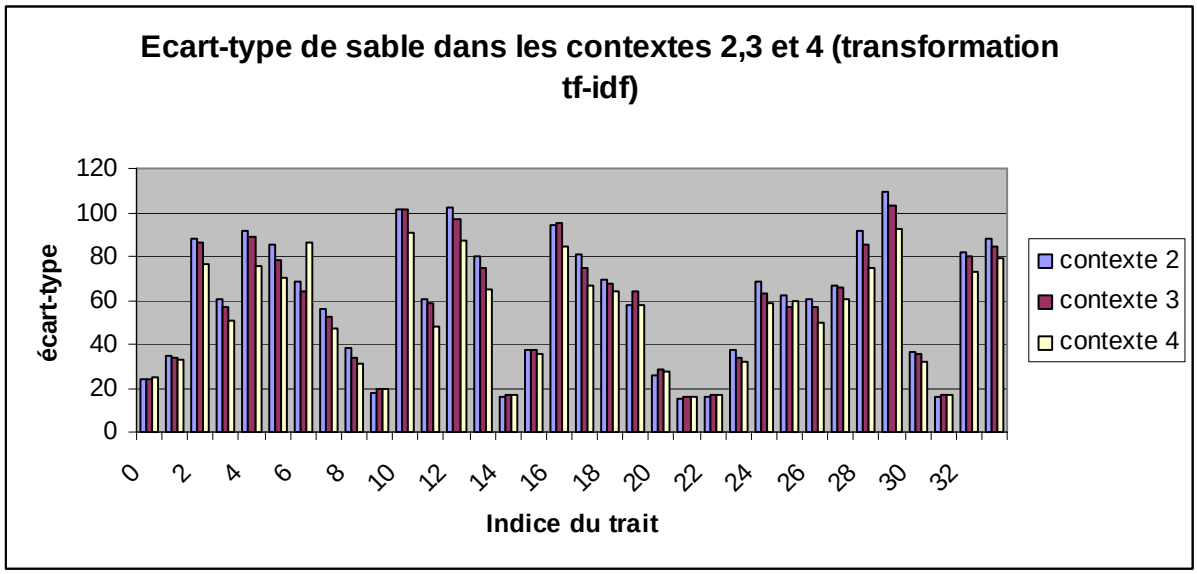

Ecart-type de sable dans les contextes 2, 3 et 4 (cooccurrences simples)
écart-type
20000
15000
10000
5000
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32
Indice du trait
contexte 2
contexte 3
contexte 4


Moyenne de sable dans les contextes 2, 3 et 4
Moyenne
300
250
200
150
100
50
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32
Indice du trait
contexte 2
contexte 3
contexte 4


Médiane de sable dans les contextes 2,3 et 4 (transformation tf-idf)
médiane
300
250
200
150
100
50
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32
indice du trait
contexte 2
contexte 3
contexte 4


Premier quartile de sable dans les contextes 2,3 et 4 (transformation tf-idf)
premier quartile
250
200
150
100
50
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32
Indice du trait
contexte 2
contexte 3
contexte 4


Troisième quartile de sable dans les contextes 2, 3 et 4 (transformation tf-idf)
Troisièma quartile
350
300
250
200
150
100
50
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32
iondice du trait
contexte 2
contexte 3
contexte 4


Ecart-type de sable dans les contextes 2,3 et 4 (transformation tf-idf)
écart-type
120
100
80
60
40
20
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32
Indice du trait
contexte 2
contexte 3
contexte 4

A6) Moyennes de traits sémantiques de rose relatives au trait /église/

Traits sémantiques de rose sélectionnés :

/vivre/,/beauté/,/cœur/,/parfum/,/amour/,/couleur/,/église/,/instrument/,/rouge/

Regroupements morphologiques : familles des traits sémantiques sélectionnés

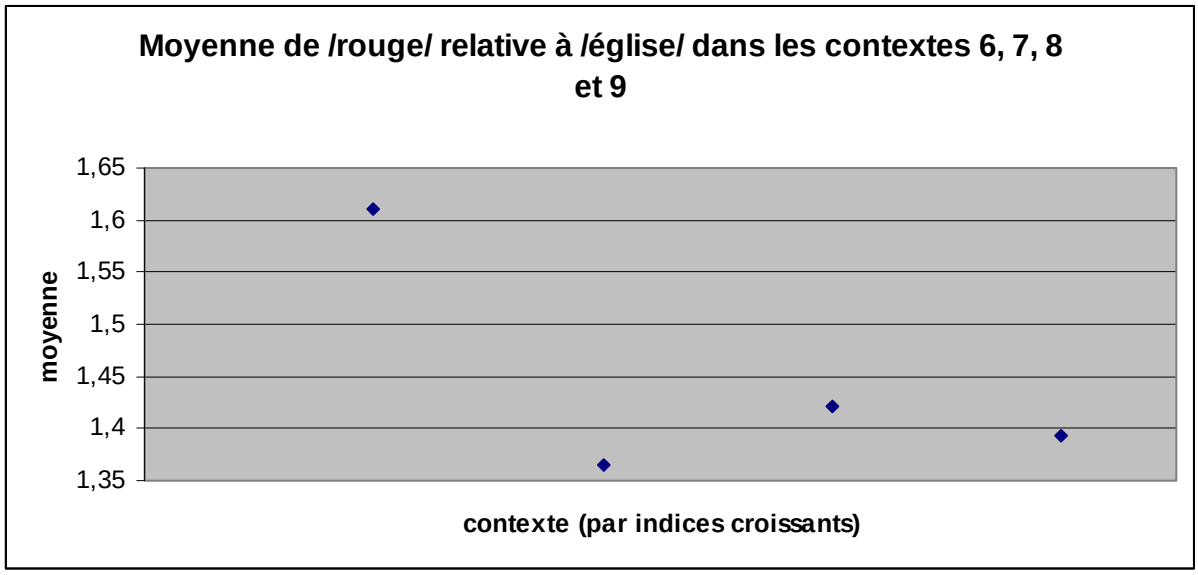
numéro	item
0	/3191/ : vivre,NOM survie,NOM survivre,VERBE vitaliste,NOM dévitalisation,NOM revitaliser,VERBE dévitalisé,ADJ dévitaliser,VERBE vitaliste,ADJ vital,ADJ vivres,NOM vitalité,NOM revitalisation,NOM vivre,VERBE survivance,NOM vitalisme,NOM vivrier,NOM vivrier,ADJ vie,NOM revivre,VERBE revitalisant,ADJ
2	beauté
4	cœur
28	parfum
52	/779/ : énamourement,NOM enamouré,ADJ amoureuse,NOM amoureux,ADJ enamourer,VERBE enamourement,NOM énamourer,VERBE amoureusement,ADV amoureux,NOM amouraché,ADJ amour,NOM
88	/2621/ : coule,NOM couleur,NOM
89	église
91	/612/ : installer,VERBE instillateur,NOM réinstallation,NOM instrumental,ADJ instrumentalité,NOM instrument,NOM instrumentalisation,NOM instrumentation,NOM installeur,NOM instrumentiste,NOM instrumentalisme,NOM instrumentaliste,ADJ instrumenter,VERBE installateur,NOM instrumentaliste,NOM instrumentaire,ADJ installeur,ADJ instillation,NOM instrumentalement,ADV installation,NOM instrumentateur,NOM instrumentaliser,VERBE instiller,VERBE réinstaller,VERBE
100	/3083/ : rouge,NOM rougeoyer,VERBE rougeot,NOM infrarouge,ADJ rougeaud,NOM rougi,ADJ rougeâtre,ADJ rougissant,ADJ rougeur,NOM rougeole,NOM rougeoiement,NOM rougeoleuse,NOM rouget,NOM rougeoleux,NOM rouge,ADJ infrarouge,NOM rougeaud,ADJ rougeot,ADJ rougeâtre,NOM rougeoyant,ADJ rougir,VERBE rougissement,NOM enrougir,VERBE rougeoleux,ADJ dérougir,VERBE

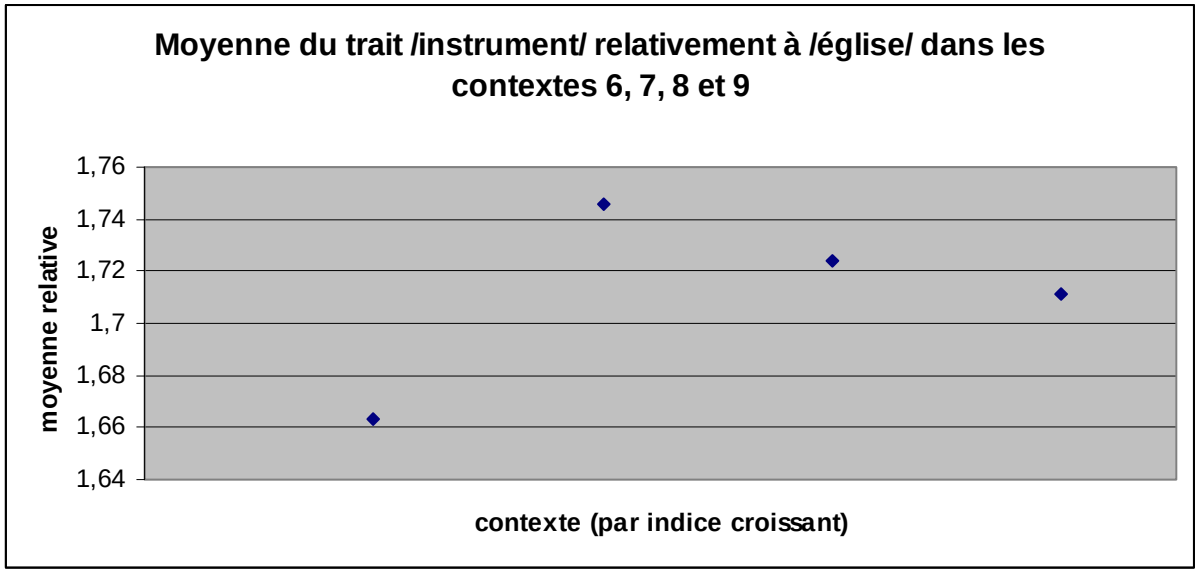
Analyse linguistique : activation de traits sémantiques de rose

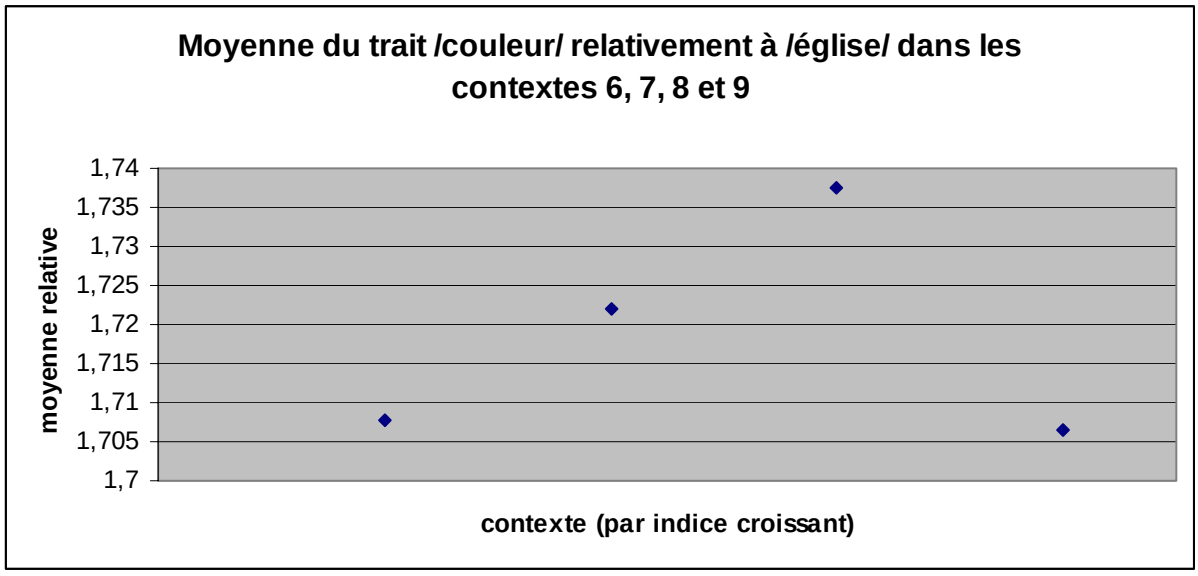
	Contexte 6	Contexte 7	Contexte 8	Contexte 9
amour			1	
beauté	1		1	
cœur			1	
couleur	1	1		
église				
instrument				1
parfum	1			
rouge	1			
vivre			1	1

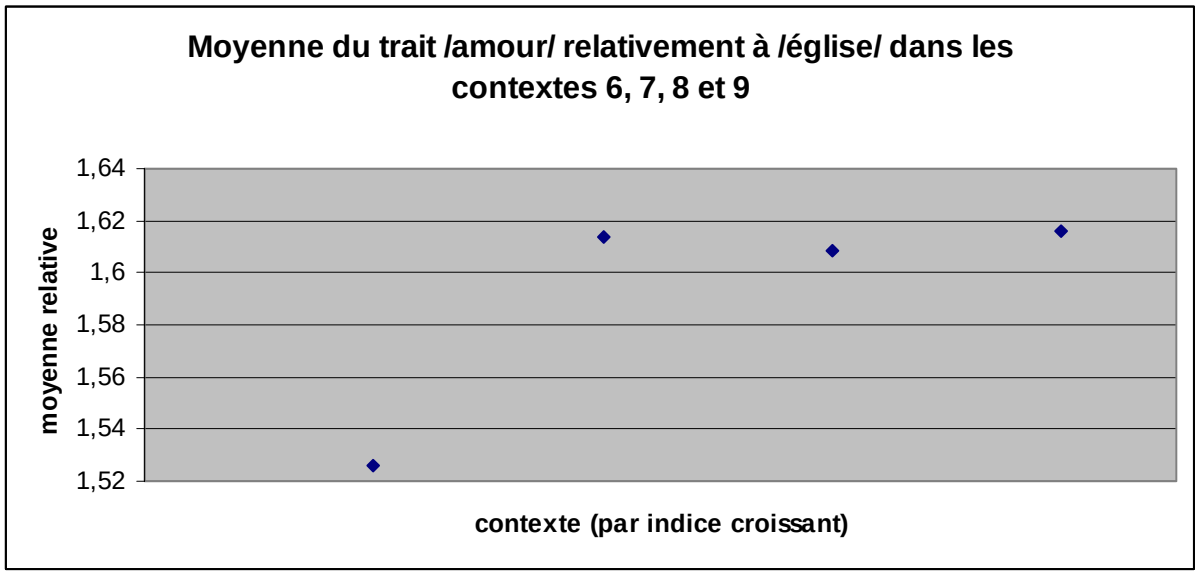
1: activation

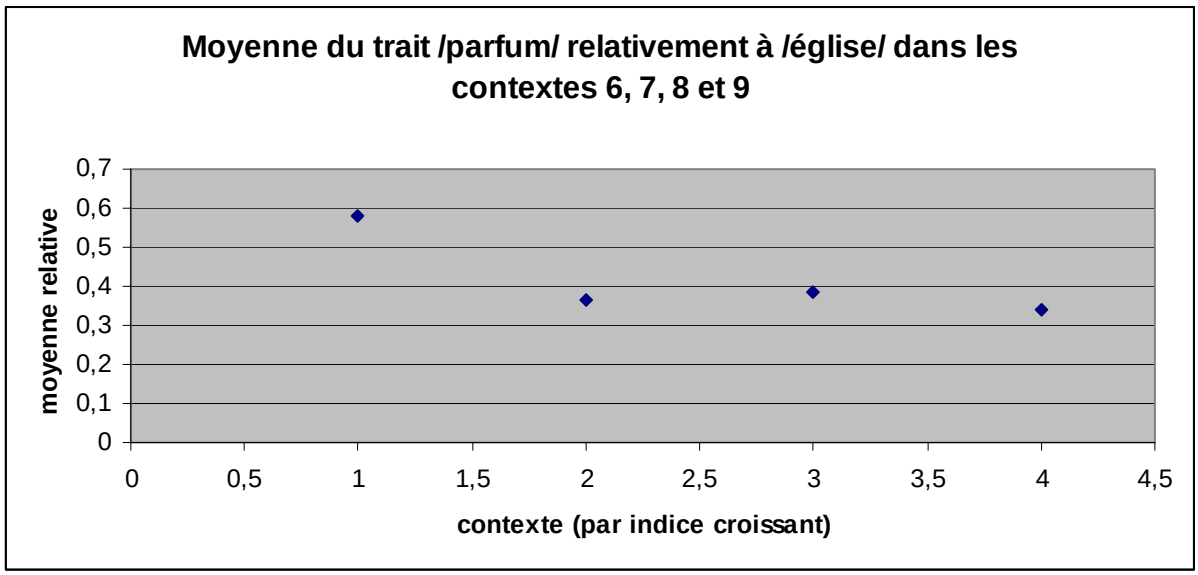
Résultats d'expériences :

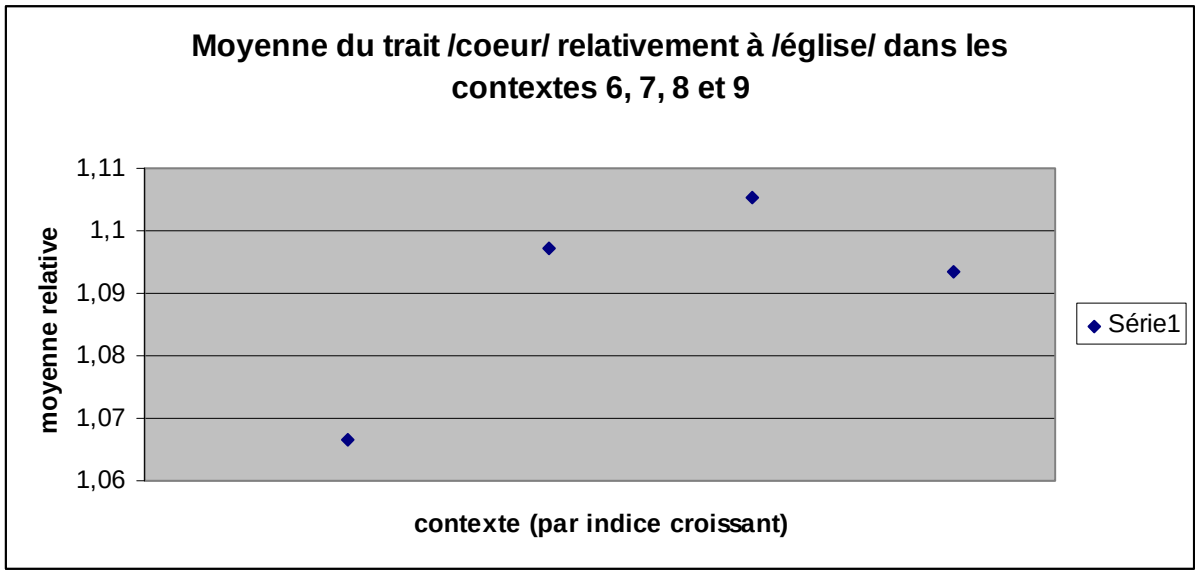

Moyenne de /rouge/ relative à /église/ dans les contextes 6, 7, 8 et 9
1,65
1,6
1,55
1,5
1,45
1,4
1,35
moyenne
contexte (par indices croissants)

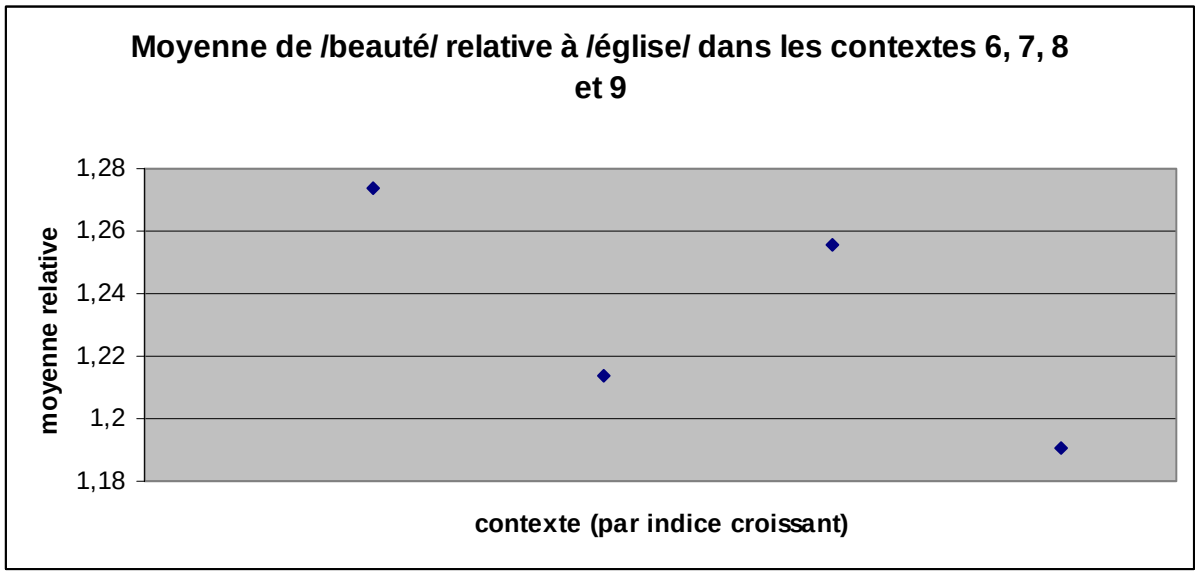

Moyenne du trait /instrument/ relativement à /église/ dans les contextes 6, 7, 8 et 9
1,76
1,74
1,72
1,7
1,68
1,66
1,64
moyenne relative
contexte (par indice croissant)

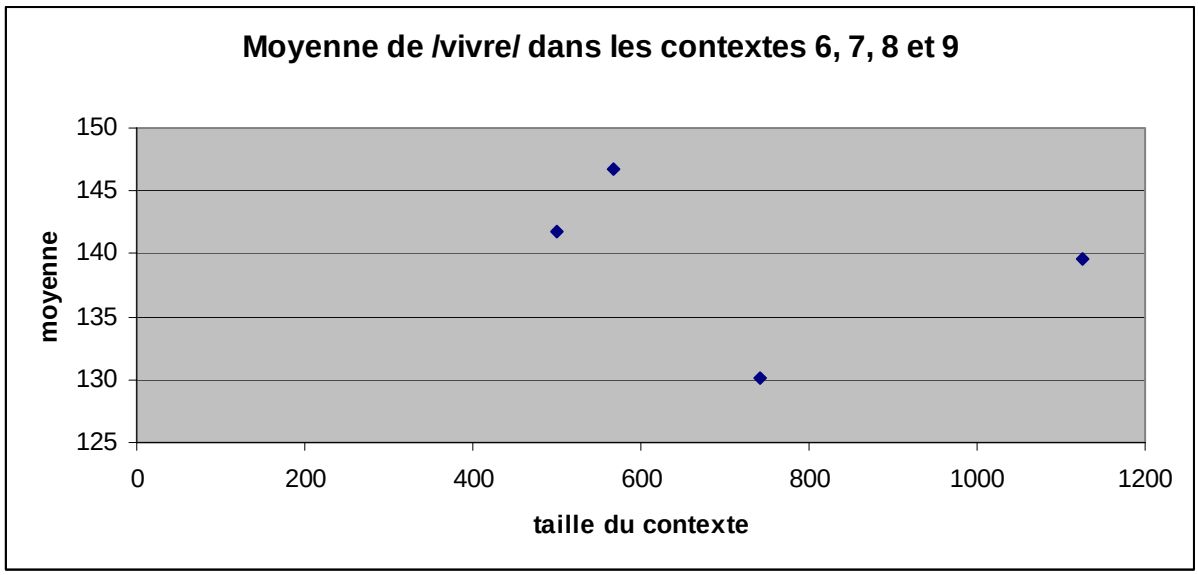

Moyenne du trait /couleur/ relativement à /église/ dans les contextes 6, 7, 8 et 9
1,74
1,735
1,73
1,725
1,72
1,715
1,71
1,705
1,7
moyenne relative
contexte (par indice croissant)


Moyenne du trait /amour/ relativement à /église/ dans les contextes 6, 7, 8 et 9
moyenne relative
1,64
1,62
1,6
1,58
1,56
1,54
1,52
contexte (par indice croissant)


Moyenne du trait /parfum/ relativement à /église/ dans les contextes 6, 7, 8 et 9
moyenne relative
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
0,5
1
1,5
2
2,5
3
3,5
4
4,5
contexte (par indice croissant)


Moyenne du trait /coeur/ relativement à /église/ dans les contextes 6, 7, 8 et 9
moyenne relative
1,11
1,1
1,09
1,08
1,07
1,06
contexte (par indice croissant)
Série1


Moyenne de /beauté/ relative à /église/ dans les contextes 6, 7, 8 et 9
moyenne relative
1,28
1,26
1,24
1,22
1,2
1,18
contexte (par indice croissant)


Moyenne de /vivre/ dans les contextes 6, 7, 8 et 9
moyenne
150
145
140
135
130
125
0
200
400
600
800
1000
1200
taille du contexte