Mondes lexicaux stabilisés et analyse statistique de discours

Max Reinert[1]

PRINTEMPS – CNRS - UMR 8085
Université de Versailles - Saint-Quentin-en-Yvelines

Abstract

The main tools used in the *"ALCESTE"* method were conceived around thirty years ago. Therefore, the presentation I will give offers a reflection on the already long tradition of textual data analysis and on the contradictions that still arise in the exploitation of results. It is also an introduction to my latest research on the analysis of *Stabilised Lexical Worlds,* and to the latest version of ALCESTE.

Résumé

Les principaux outils de la méthode « *ALCESTE* » ont été conçus, il y a environ une trentaine d'années. Cette communication propose une réflexion sur une pratique de l'analyse des données textuelles déjà bien ancienne, du fait des contradictions qu'elle suscite *toujours aujourd'hui* dans l'exploitation des résultats. Elle introduit aux derniers travaux de son auteur sur l'analyse des *mondes lexicaux stabilisés* et à la dernière version de cette méthodologie.

Mots-clés : méthode ALCESTE, analyse de discours, mondes lexicaux stabilisés, postures énonciatives, Peirce.

1. Introduction

Les méthodes statistiques d'*analyse exploratoire* des données telles que les a introduites J.-P. Benzécri (1973) à la fin des années soixante ont montré, depuis bien longtemps, leur efficacité dans le domaine des statistiques textuelles (notamment et surtout son analyse factorielle des correspondances). Pourtant une question reste en suspens : Qu'est-ce que de telles approches exploratoires peuvent représenter d'un discours ? C'est en tout cas la question que nous ne cessons pas de nous poser dans notre exploration des discours avec la « méthode *ALCESTE*[2] ». Par exemple, pourquoi cette dernière est souvent utilisée pour préparer une analyse de contenu, alors que son algorithme ignore tout des structurations syntaxiques des énoncés, peut même ignorer la séparation de ces énoncés entre eux — en tant que parties, paragraphes ou phrases — sans grandes conséquences sur les résultats ? Nous avons expérimenté maintes fois par nous-même ces jeux fluctuants entre mondes lexicaux et contenus, et nous aimerions, aujourd'hui, présenter nos conceptions pour comprendre leur lien avec l'algorithmique adoptée, d'autant qu'est apparu rapidement un autre phénomène : celui d'une stabilisation des mondes lexicaux.

[1] max.reinert@printemps.uvsq.fr.

[2] Sigle pour « Analyse des Lexèmes Cooccurrents dans un Ensemble de Segmentations du Texte Etudié ».

Rappelons le dispositif d'analyse, simplifié à l'essentiel : Il consiste à découper le texte étudié en segments de longueur comparable (généralement entre 15 et 100 occurrences), pourvu que leur nombre ne dépasse pas 40 000, et à étudier la variation des distributions des mots pleins dans ces segments (3 000 mots pleins *analysés* maximum).

Ce modèle simplifié de représentation statistique d'un discours suffit à mettre en évidence, du moins dans l'analyse de certains corpus, une tendance du vocabulaire à se distribuer dans *des mondes lexicaux stabilisés*[3]... Il y a donc une sorte de bizarrerie qui consiste à retrouver des mondes lexicaux se ressemblant entre des analyses portant sur des corpus très différents, quant à leur contenu, quant à leur taille, et même quant à leurs conditions de production[4], alors même que les utilisateurs (et moi-même) découvrions par ailleurs, à travers lesdits résultats des analyses particulières, les contenus de nos expériences de lecteurs des textes étudiés. Il a bien fallu concevoir ces faits comme n'étant pas contradictoires. Cela nous a conduit à rechercher les contenus ailleurs que dans les textes… Nous avons dû quitter l'approche des textes, conçue comme supportant des représentations, pour une approche des textes en termes de traces d'activité, traces de discours possible. Nous ne disons pas que ces traces ne permettent pas de se représenter, mais elles « représentent » aussi comme les traces de toute activité humaine.

Nous désirons d'abord présenter un bref résumé de notre conception de la relation entre *contenus* et *mondes lexicaux,* afin de comprendre ensuite l'enjeu de l'outil nouveau de la version 5 pour mesurer la stabilisation des mondes lexicaux, à travers l'exemple notamment de l'analyse d'un dictionnaire[5].

2 La problématique « ALCESTE »[6]

La conception du discours soutenant notre approche n'en retient qu'un tout petit aspect pouvant concerner toute activité, ce que souligne d'ailleurs déjà l'origine de ce mot, « discours » venant d'un mot latin « discurere » signifiant « aller de-ci de-là », et désignant donc un type d'activité indépendant du langage ! Le point de vue que nous soutiendrons affirme d'abord le discours comme activité sociale. Notre collaboration avec le sociolinguiste Pierre Achard peut être à cette occasion rappelée, lui qui appelait *discours*, *« l'usage du langage en situation pratique, envisagé comme acte effectif, et en relation avec l'ensemble des actes (langagiers ou non) dont il fait partie »* (1993, p10). Et cela a eu des conséquences sur notre conception du contenu, le contenu étant d'abord le contenu d'une activité. Prenons l'exemple d'un chasseur allant par monts et par vaux : il distingue des traces là où un promeneur ne voit rien ! Ces traces sont liées au contenu de son expérience. Le contenu est premier, et les traces sont secondes : elles ne sont visibles qu'en rapport avec un contenu qui leur est antérieur. Le contenu est premier par effet de la mimesis, c'est-à-dire des habitudes de ce chasseur intégrées aux habitus liés de son milieu de vie ; les traces sont secondes et elles le questionnent. Les traces problématisent le contenu, et, dans un troisième temps, notre chasseur va chercher à les intégrer à une représentation plus large. Ce qui cause le contenu,

[3] Ce qu'Etienne Brunet semble également avoir retrouvé en utilisant une méthodologie voisine (Brunet, 2006)

[4] Voir le §3.1 : « Aurélia » est l'œuvre d'un auteur, alors que la revue surréaliste SaSdlR est l'œuvre d'un groupe…

[5] Version antérieure à 1997 du petit Larousse d'environ 16 millions de caractères de texte.

[6] Se reporter également à : « Postures énonciatives et mondes lexicaux stabilisés en analyse statistique de discours », (Reinert, 2007).

lui, reste caché, car il se fond avec une origine de ce qui anime (la mimésis), alors qu'on ne peut en saisir que des signes épars ne prenant vraiment sens que dans l'après-coup (troisième temps) d'une représentation médiate. Ainsi le contenu est davantage dépendant d'une *matérialité* des pratiques que des *formes* construites. Il dépend des usages du monde en actes, ce qui ne nécessite nullement qu'il soit représenté (du moins consciemment ou médiatement).

Dans l'activité langagière, les mots pleins constituent, selon nos hypothèses, des traces *possibles* des contenus de nos activités. Ils ne sont pas *les* signifiants mais bien *des* traces possibles de ce contenu en acte. Bien sûr, il existe des dictionnaires, dont on ne doit pas oublier qu'ils ne signifient rien sans un usage. Si les mots appartiennent à un bien commun, c'est d'abord par leur matérialité, c'est-à-dire dans la mesure où ils s'intègrent aux activités et usages d'une époque, d'un lieu, d'une population ; ils ne sont jamais seuls, aussi les contenus de ces usages impliquent-ils toujours des constellations de mots et de marques de toutes sortes. Les mots pleins ont cependant la faculté de susciter des contenus en tant qu'ils stabilisent nos visions du monde, *mais le contenu n'est pas dans le mot ; il est dans l'acte, dont le mot est une trace.* Même en lecture, le contenu est premier : c'est par lui que notre intérêt de lecteur peut se porter sur la différenciation des mots, puis leur combinaison. Ainsi la trace reste seconde et si elle dépend des codes présents sous les yeux du lecteur, elle dépend également de son attente, de ce qui fait sens pour lui à ce moment précis de la lecture. Et cela peut varier d'une lecture à l'autre, d'un lecteur à l'autre, aussi la constellation des mots « pleins » – c.a.d. concernés par le contenu de l'acte – est fluctuante. C'est la raison qui nous a conduit à une procédure de découpage du texte à la fois fruste et variable pour ne repérer des traces potentielles de ce contenu qu'à travers la cooccurrence des mots pleins. *Notons que cette cooccurrence ne peut être mise en évidence sans un acte de lecture ou d'énonciation permettant de la détacher comme trace*… c'est-à-dire sans la lier à un *contenu* ! Ce détachement, un algorithme sans mémoire ne peut pas l'obtenir sans un artifice : il consiste, avec *ALCESTE*, en une segmentation arbitraire du texte[7]. Mais ce découpage permet, une fois institué, de calculer des probabilités pour ces cooccurrences, et donc de relever dans l'après-coup, ce qui peut se répéter et souligner une pertinence, un mode d'être.

Cela étant dit, il y a un second problème pour l'estimation statistique de ces liens… C'est la très faible probabilité de leur réalisation. Prenons l'exemple de l'analyse du corpus d'étalonnage d'environ 26 millions de caractères pour constituer la version 5 du logiciel. Son analyse conduit à un tableau de données comprenant près de 40 000 unités de contexte en lignes et 3 000 mots pleins en colonnes, le nombre de « 1 » retenus étant d'environ un million, soit en proportion, moins de 1 pour 100… Imaginons maintenant cet espace de représentation et les distances deux à deux entre unités de contexte (U.C.) avec ses près de 3 000 dimensions… Il suffit de changer un peu d'arbitraire dans le découpage des unités de contexte (U.C.) pour que des centaines de dimensions disparaissent, pour que des centaines de nouvelles dimensions apparaissent, introduisant des distorsions considérables dans les distances entre U.C. prises deux à deux… Cet espace global ne peut servir de référentiel exact pour une description. L'analyse factorielle des correspondances ne sert pas vraiment à « résumer l'information d'un nuage de point » sur un axe au prix d'une certaine déformation. Elle sert surtout dans le cas considéré *à construire des dimensions possibles pour un espace de représentation relativement stabilisé.* C'est la raison pour laquelle nous avions opté à

[7] Si du point de vue théorique l'aspect arbitraire est essentiel, du point de vue pratique, l'algorithme tient compte (en partie) de la ponctuation, ponctuation qui peut donc être utilisée par un analyste pour marquer le texte en fonction de sa propre lecture.

l'origine de cette recherche, pour une classification descendante, les deux premières classes se distinguant à partir du premier axe d'une A.F.C.…

Revenons aux résultats : Non seulement les résultats sont relativement stables entre les différents essais d'analyse d'un corpus donné, mais il se trouve qu'ils peuvent l'être entre analyse de corpus différents ! Une telle méthode ne permettrait pas de constituer des mondes lexicaux ainsi stabilisés *si un autre processus n'entrait pas également en interaction, car il a bien fallu admettre l'existence d'un mode de fabrication de la différence relativement indépendant des domaines de connaissance dont relevaient ces corpus !* Aussi la mimésis et le jeu des contenus mis en œuvre dans les interprétations ne peuvent expliquer à eux seuls ce processus… *ou plus précisément ils ne peuvent prendre sens qu'en tant que leur formation même est dépendante d'un processus maître,* commun à toute activité discursive, à l'origine des rythmes et scansions observés dans toute énonciation ou acte de lecture.

Nos hypothèses sur ce processus rejoignent pour une bonne part celles de Peirce[8], la principale étant que l'activité discursive, comme toute activité humaine (donc sémiotique), s'introduit à travers un certain ordre. La sémiose, en tant qu'elle concerne une conscience, part de signes d'abord perçus immédiatement — notion de fondement du representamen[9] chez Peirce, de contenu pour nous — pour se refermer sur des signes construits médiatement (l'interprétant chez Peirce), en passant par une période plus ou moins longue exprimant le temps de l'expérience, où l'objet du signe se montre (à travers ses traces) en résistant au contenu immédiat, obligeant ainsi à construire une représentation médiate ou concept. Rappelons que phénoménologiquement ces différents moments de la sémiose se distinguent pour Peirce par des modes d'être différents : priméité, secondéité et tiercéité. Pour nous, ces modes d'être vont constituer dualement des registres de la réalité, accessibles à travers des postures. Par exemple, la priméité, *le un*, est dans ce qui peut être saisi en soi, par exemple une sensation, une couleur, une odeur. Elle est liée à la possibilité d'un quelque chose. Elle est liée également selon nous à une posture particulière permettant de rendre compte de ce monde de la priméité : Celui-ci n'est « en soi » que par le fait qu'on le voit seulement. C'est-à-dire, sans se voir le voir et sans se savoir vu soi-même ! Cette posture première, on l'appellera, en reprenant les termes de Achard[10], la *posture du témoin*. Elle introduit à un monde perçu en soi, comme une image ; C'est la raison qui nous fait associer, avec d'autres auteurs[11], la priméité au registre de l'Imaginaire. C'est un thème classique de philosophie que de démontrer l'impossibilité de partir des sensations pour baser avec certitude l'existence du monde[12]. Pour Peirce, la priméité n'exprime en fait qu'une simple *possibilité* des choses (et non pas leur existence réelle).

La secondéité, pour Peirce, se constitue de ce dont on fait l'expérience, ici et maintenant, quand il y a résistance à la simple apparence des choses. Autre chose résiste à ce que l'on a

[8] Voir notamment « *Ecrits sur le signe »,* assemblés, traduits et présentés par G. Deledalle (Peirce, 1978).

[9] Le « representamen » correspondrait plutôt à ce que nous nommons « trace » dans cet exposé.

[10] Pierre Achard introduit les « postures » ou « positions » d'énonciation — témoin, acteur et patient — dans le cadre particulier d'une analyse d'un questionnaire posé à d'anciens appelés de la guerre d'Algérie : « Une approche discursive des questionnaires : l'exemple d'une enquête pendant la guerre d'Algérie » (Achard, 1991, p 21 et s.). Nous reprenons cette terminologie en l'étendant au cadre qui est le nôtre.

[11] Notamment le processus interprétatif (Everaert-Desmedt,1990). Voir également Des fondements sémiotiques de la psychanalyse (Balat, 2000).

[12] Voir par exemple Descartes et son *Discours de la méthode*… d'où il déduit que la seule certitude qu'il peut avoir d'une existence est dans le fait qu'il pense…

cru voir. Altérité, opposition, travail, combat, épreuve, expérience en acte, telles sont les expressions de ce second mode d'être, qui implique à la fois la conscience d'une expérience passée, et la conscience d'un manque actuel qui ouvre à l'engagement. Au travers des traces, par exemple, ce conflit s'exprime justement comme vacillement entre *contenu* d'un habitus et *sens* dépendant des aléas d'une expérience actuelle. La posture que nous lui associons est celle de *l'acteur*[13] dont l'existence exprime ce pour quoi il s'engage.

Enfin, avec la tiercéité s'introduit l'idée d'un troisième terme nécessaire pour permettre une issue (temporaire) à une lutte sinon sans issue. Le symbole est ce par quoi ce qui divisait devient lien pour l'avenir. Après un long conflit, un mauvais compromis peut ouvrir à de nouvelles formes de vie, et apparaître bénéfique dans l'après-coup. La mémoire et le présent ne se rencontrant pas toujours, il faut trouver un mode d'aménagement : par exemple, si l'introduction d'une porte vitrée dans un parcours habituel conduit à quelques heurts, il peut suffire d'y coller une étiquette, pour aider à modifier ses habitudes à cet endroit (sans attendre donc que la porte ou la tête finisse par céder !). L'étiquette est ce qui permet de donner une nouvelle forme à l'activité future… Collectivement on ne cesse pas de s'inventer des signes qui permettent de transformer nos habitudes dans de nouveaux concepts de manière à ce qu'elles soient compatibles avec un réel qui reviendrait sinon en catastrophe ! Ce registre du symbolique, en introduisant un troisième terme (le signe), ouvre à *la posture du patient* pour reprendre la terminologie de Achard, elle ouvre au raisonnement, à la prévision… à condition de prendre le temps nécessaire pour comprendre !

Les catégories peirciennes de la *priméité, secondéité* et *tiercéité* ne sont pas simplement phénoménologiques, elles expriment également une temporalité dans l'émergence de nouveaux signes (sémiose), rythmant de différents moments les activités par des changements constants de postures dans la manière d'appréhender les choses.

Ces idées s'étant imposées à nous à partir de l'expérience des mondes lexicaux, il est temps d'y revenir pour tenter d'en pister maintenant les traces !

3. Les expériences sur les mondes lexicaux stabilisés

3.1 L'exemple d'un air de famille entre deux analyses « ALCESTE »

Le fait d'avoir été sensible à travers des « analyses » diverses à des « airs de famille » nous a conduit à une première expérience pour tenter de cerner ces « ressemblances ». La comparaison portait sur les analyses de deux corpus littéraires : une revue sur le surréalisme « le surréalisme au service de la révolution » et une œuvre de Nerval « Aurélia »[14]. Elle a consisté en deux analyses « *ALCESTE* » indépendantes (ancienne version), le nombre de classes maximum demandé étant limité à 3. Le vocabulaire spécifique de chaque classe est présenté à l'aide d'une liste de mots ordonnés par $\chi 2$ décroissant (cf. Tableau 1).

[13] Terme introduit par Achard (o.c.)

[14] Voir « Les 'Mondes lexicaux' des six numéros de la revue 'Le Surréalisme au Service de la Révolution' » (Reinert, 1997).

	Les trois premiers mondes lexicaux de l'analyse d'AURELIA	Les trois premiers mondes lexicaux de l'analyse des 6 numéros du SaSdlR
1	vis, fleur+, yeux, figure+, lumiere+, soleil+, terre+, etoile+, mur+, vaste+, corps, couleur+, rayon+, brill+er, devenir., elev+er, entour+er, sembl+er(40), nuag+e(9), vieill<, cru+, eclaire+, grand+, profond+, teinte+, blan+14, air+, feu+, lune+, nuit+;	noir+, rouge<, eau+, femme+, main+, nuit+, pied+, soleil, statue+, terre+, tete+, nu+, blan+14, air+, cheminee+, fille+, yeux, vill+2, mur+, rose+, bleu+, vert+, arbre+, bouche+, boule+, bras, cheveu+, couleur+, dent+, droite+, feu+, soir+, ventre+;
2	paris, ami+, amour+, cimetiere+, faute+, lettre+, papier+, pus, tombe+, donn+er, laiss+er, regard+er, rendre., suivre., visit+er, aller., bague+, larme+, os+er, retrouv+er, vierge+, eprouv+er, poe+19, malade+, maison+, parent+, veille+, chant+er, cherch+er, pass+er, rappel+er, pauvre+, fou+ ;	allemand+, francais+, milita+ire, ouvrier<, decembre+, article+, camarade+, commissariat+, communis+me, ecole+, ecrivain+, fascis+me, genera+l, guerre+, journa+l, lettre+, livre+, paix, police+, revue+, sport+, titre+, adress+er, ecrire., publi+er, bourgeois<, collaborat+ion, edit+ion ;
3	dieu+, esprit+, vie+, monde+, double+, evenement+, humain+, vrai+, etude+, idee+, science+, exist+er, amer+, chretien+, vague+, maladie+, rapport+, souvenir+, comprendre., occup+er, vivre., entier+, mort+, perdu, oncle+, desespoir+, fois, proie+, raison+, sommeil ;	concret+, evoluti+f, expressi+f, humain+, mora+l, particulier+, pratique+, connais-sance+, contenu+, developpement+, devenir+, dialectique, domaine+, element+, esprit+, homme+, humour+, idee+, image+, materia-lis+me, methode+, monde+, mot+, moyen+, nature, objet+, phenomene+, processus ;

Tableau 1[15] : Comparaison des trois « mondes lexicaux » des analyses « Aurélia » et « SaSdlR »[16]

Ligne 1 (cf. Tab. 1) : le vocabulaire de la priméité est très ressemblant entre les deux analyses. Ce vocabulaire comme registre de l'Imaginaire est particulièrement adéquat ici, puisque les unités de contexte de l'œuvre de Nerval relevant de la classe 1 recouvrent en bonne part les retranscriptions de rêves, et dans l'autre oeuvre, les unités de contexte proviennent plus particulièrement des poésies et jeux d'associations (style plutôt descriptif : posture de témoin).

Ligne 2 (cf. Tab. 1) : L'aspect conflictuel de la secondéité s'exprime dans « Aurélia » par l'expression contrastée d'affects ; Dans la revue surréaliste, le conflit est plus directement exprimé en termes d'engagement politique. Le Réel ne peut véritablement constituer un registre n'étant pas directement exprimable. A travers cette comparaison d'analyses s'exprime également une double différence de tonalité expressive entre *ce qui est éprouvé* avec Nerval et *ce qui est vécu comme engagement* avec le groupe surréaliste… Aussi l'expression de la secondéité semble dépendre de plusieurs modes de restitution de l'expérience, soit orienter vers les états d'âme, les passions, ou soit orienter vers les objectifs mondains de l'engagement. Quel qu'en soit le mode, cette sorte de dimension (le Réel) est *indiquée*[17] à travers un style plutôt narratif, incluant une temporalité, un mouvement, une lutte, une fracture, mettant en exergue la dimension du sujet (posture de l'acteur).

Ligne 3 (cf. Tab. 1) : Le vocabulaire de la tiercéité implique des mots abstraits en rapport avec, ce qui, dans une société, permet une médiation des conflits : lois, institutions, justice,

[15] Ce tableau est publié page 394 (Reinert, 2001).

[16] Les 6 numéros de la revue d'André Breton, « le Surréalisme au Service de la Révolution », parue entre 1930 et 1933. Résumé d'une analyse avec la version 2 du logiciel (Reinert, 1997, note 1).

[17] Cette indexicalité est perceptible au niveau des mondes lexicaux par le grand nombre de noms propres.

connaissance par concepts, religion, pouvoir politique, etc. Tous ces mots à la fois *représentent et organisent* des formes de vie éventuellement en donnant un sens social à la souffrance et à la frustration. Le registre qui en dépend est clairement celui du Symbolique. Le vocabulaire associé est caractérisé par son niveau élevé d'abstraction. Dans les deux œuvres présentées, ce vocabulaire est caractéristique des parties où l(es)'auteur(s) cherche(nt) une solution aux contradictions qui l(es)'assaillent : retrouver des êtres chers décédés pour Nerval, concilier engagement politique et création poétique, pour le groupe de Breton. La recherche de raisons est un moyen de supporter cette tension du réel à travers de nouveaux symboles permettant une espérance pour l'avenir, tout en facilitant l'adoption de nouvelles formes de vie.

3.2 L'étalonnage des mondes lexicaux stabilisés dans la version 5 du logiciel

Les « airs de famille » entre mondes lexicaux obtenus dans des analyses différentes sont trop évanescents pour en rendre compte en dehors d'interprétations subjectives. Comment mesurer cette stabilisation ? Le paragraphe précédent rend déjà sensible le fait que le vocabulaire des deux registres, imaginaire et symbolique, semble plus stabilisé que le vocabulaire indexant du réel. Une première tentative (vers 1995-96) pour construire des registres de vocabulaire à partir d'une analyse de contenu avait abouti à distinguer non pas à 3 registres mais 4[18]. Il fut cependant impossible de la publier du fait de la subjectivité de l'analyse. Aussi nous avons décidé de mettre au point un outil pour étudier plus systématiquement cette stabilisation des mondes lexicaux, d'abord en augmentant les capacités d'analyse du logiciel afin de traiter des corpus de grandes tailles, et ensuite de manière à construire un corpus d'étalonnage permettant un classement statistique automatisé d'un vocabulaire le plus vaste possible.

Le corpus d'étalonnage que nous avons constitué dépend bien sûr également des arbitraires d'une époque, très concrètement ici, des possibilités d'accès aux œuvres sur le Web. Nous l'avons composé essentiellement entre 1998 et 2003. Il réunit aujourd'hui une cinquantaine d'œuvres, composé pour un tiers d'œuvres du XIX^{e} siècle pour un tiers d'œuvres du XX^{e} siècle, et pour un tiers d'œuvres antérieures au XIX^{e} siècle. Sa taille est d'environ 26 millions de caractères[19] :

---

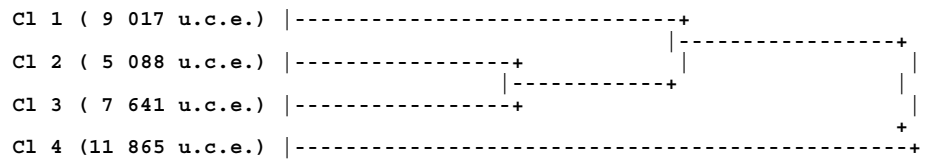
[18] Nous avions tenté une analyse de contenu sur plusieurs milliers de mots pleins, classement abandonné par la suite, qui a été cependant utile pour organiser une première vision des mondes lexicaux stabilisés en mettant sur la voie d'une distinction de deux mondes lexicaux pour appréhender le registre du vocabulaire de la secondéité.

[19] Nombre d'œuvres utilisées sont issues du site de l'association des bibliophiles Universels (*http://abu.cnam.fr/)* que je remercie particulièrement. D'autres viennent d'autres sites dont j'ai perdu la trace, certaines, indirectement de l'INALF, et de retranscriptions personnelles par des chercheurs ou des étudiants, d'autres m'ont été offertes comme le corpus sur les discours du Général de Gaulle (D. Labbé). Je profite de cette occasion pour remercier tous ceux avec qui j'ai travaillé, et qui m'ont aidé dans cette recherche, ne serait-ce que par leur amical soutien. Voici la liste de ces oeuvres : *Les faits* et *l'avenir dure longtemps* de Louis Althusser *;* Le *café bleu* d'Edmond Amran el Maleh *; Le colonel Chabert, les Illusions perdues,* d' Honoré de Balzac*, les dissertations sur l'érotisme*, de George Bataille *; Les Fleurs du Mal, Mon cœur mis à nu,* de Charles Baudelaire ; *Les deux sources de la morale et de la religion,* d'Henri Bergson ; *Le diable amoureux* de Jacques Cazotte ; *Voyage au bout de la nuit*, Ferdinand Céline ; *Les chants de Maldoror*, d'Isidore Ducasse ; *Différence et Répétition,* de Gilles Deleuze ; *Le discours de la méthode, Les méditations,* de René Descartes ; *La théorie physique, son objet, sa structure* de Pierre Duhem ; *Le carnet noir* (extrait traduit) de Lawrence Durrell ; *Entretiens sur la pluralité des mondes* de Bernard Fontenelle ; *Psychopathologie de la vie quotidienne* de Sigmund Freud ; *Dominique* de Eugène Fromentin ; *Le capitaine Fracasse* de Théophile Gauthier ; Les discours du Général de Gaulle, 79 allocutions ou conférences de presse *; L'année terrible (1871), L'art d'être grand-père (1877), Bug-Jargal (1825), Les contemplations, La légende des siècles, (1859), Notre Dame de Paris, Quatre-vingt-treize, Les Chansons des rues et des bois, La fin de Satan,* de Victor Hugo ; *Les preuves (1898),* de Jean Jaurès ; *Fonction et champ de la parole et du langage en psychanalyse*, de Jacques Lacan (Ecrits, pages 237-

L'analyse effectuée avec la version 5 est une analyse simple sur les U.C.E.[20], limitée à la partition en 4 classes. Elle a permis de classer près de 10 000 formes réduites dans 4 mondes lexicaux stabilisés[21]. Voici un extrait des caractéristiques de l'analyse :

Nombre de caractères	26 389 562
Nombre d'occurrences	4 472 007
Nombre de formes distinctes	92 156
Nombre de formes réduites analysées	2 979
Nombre d'occurrences par U.C.E	127
Nombre de formes supplémentaires	6 819
Nombre d'unités de contexte élémentaires (U.C.E.)	33 821
Nombre d'occurrences analysées	1 089 926
Poids du tableau effectivement analysé par la C.D.H.	998 885
Nombre d'U.C.E. effectivement classées par la C.D.H.[22]	33 611 soit 99.38%

L'arbre de la C.D.H. (obtenu par analyse simple en 4 classes) :



322) ; *Les liaisons dangereuses*, Pierre Choderlos de Laclos ; *La condition humaine*, André Malraux *; Manifeste du parti communiste,* Karl Marx & Frédéric Engels ; *Aurélia,* Gérard de Nerval ; *Le Réel,* de Daniel Parrocchia ; *Les pensées*, Blaise Pascal ; *Des Opinions des philosophes*, Plutarque (46-120) ; *Anonymes*[19]*,* 1997-1999 *; La science et l'hypothèse,* Henri Poincaré ; *Le diable au corps,* Raymond Radiguet (1903-1923) *; Souvenirs d'enfance et de jeunesse*, Ernest Renan ; *Les confessions,* Jean-Jacques Rousseau ; *Cours de linguistique générale,* Ferdinand de Saussure ; *La Révolution Surréaliste* (12 numéros), *Le Surréalisme au Service de la Révolution* (6 numéros entre 1929-33) ; *Instants de vie*, Virginia Woolf ; *Voyages en France pendant les années 1787, 1788, 1789*, Arthur YOUNG ; *La curée,* Emile Zola.

[20] Analyse simple sur les unités de contexte élémentaires. Ce choix se justifie par le fait que l'analyse de grands corpus est insensible au changement local de grandeur des unités de contexte… incitant d'ailleurs fortement à penser à l'aspect fractal de la structure I.R.S. mise en évidence.

[21] Son analyse permet de distinguer au moins 6 classes, chaque posture de l'activité énonciative pouvant être associé non pas un monde lexical mais à deux, et correspondant aux 6 premières classes obtenues lors de l'analyse du corpus d'étalonnage. Mais nous ne présenterons ici que les 4 premières classes, celles-ci ayant servies à construire les 4 clés topiques de la version 5.

[22] A chaque pas de la classification, certaines U.C.E. peuvent ne pas être classées ; d'autre part (ce n'est pas le cas de cette analyse, mais de la suivante) certaines petites classes peuvent se trouver éliminées. Ce pourcentage informe de la perte subie globalement pour obtenir le classement proposé.

Clés topiques	Traces des quatre mondes lexicaux stabilisés
Clé A Imaginaire (classe 1)	ombre+, noir+, ciel+, nuit+, sombre+, terre, soleil+, vent, fleur+, front+, blanc+, oiseau+, pied+, eau, aile+, oeil+, bois, tete+, arbre+, etoile+, bleu+, mer+, bruit+, mur+, pierre+, cieux, lueur+, souffle+, flot+, lumiere+, horizon+, gouffre+, rose+, sang+, rouge+, vert+, aube+, nuage+, herbe+, flamme+, azur+, aurore+, pale+, foret+, ame+, feu, cheveu+, abime+, toit+, tombe+, bras, astre+, tenebre+, bouche+, fenetre+, obscur+, clarte+, mont+, ange+, eclair+, regard+, doigt+, dent+, voix, hideu+x, fumee+, immense+, brume+, mort+, parfum+, vague+, neige+, lune+, branche+, air, spectre+, reve+, sinistre+, deuil+, plein+, rayons, plaine+, rayon, bas, fer, tombeau+, chair+, voile+,
Clé B Réel -affect (classe 2)	sentiment+, amour+, amitie+, bonheur+, plaisir+, coeur+, amie+, aimer, aime, heureu+x, tendre+, sentais, malheureu+x, peine+, senti+, cruel+, chagrin+, aimable+, ami+, voyais, malheur+, cher+, crainte+, consolation+, soin+, plaindre, vivre, tort+, dis, courage+, charme+, faiblesse+, ecrire, passion, gout+, rendre, regret+, croyez, aimais, maman+, croyais, souffrir, espere, sentir, digne+, promettre., fis, tendresse+, attachement, douceur+, disais, fach+er, confiance, faute+, craindre., desir+, projet+, craindre, douce+, sincere+, savais, reproche+, sacrifice+, mere+, talent+, honte+, penible+, jouissance+, croire., orgueil+, aim+er, conduite+, su, revoir, aimee, desespoir+, trouvais,
Clé C Réel -socius (classe 3)	repondit, fit, maison+, demanda, pere+, reprit, diner, duc, ecria, chambre+, jeune+, hotel+, affaire+, marquis+, argent+, heure+, prit, mari+, garcon+, mois, soir+, aller, fille+, allait, savait, trouva, rue+, comte+, compagnie+, table+, vint, dame+, jeune-homme, appartement+, pay+er, mit, journal+, bonne+, salon+, piece+, baron, trouvait, donna, venait, theatre+, visite+, cabinet+, gentilhomme, alla, pauvre+, client+, ami+, libraire+, joli+, faisait, disait, famille+, boulevard+, matin+, actrice+, province+, comtesse+, conversation+, voiture+, billet+, souper, payer, directeur+, boutique+, drole+, fortune+, cour, abbe+, quartier+, ajouta, acheter, soiree+, camarade+, journaliste+, frere+,
Clé D Symbolique (classe 4)	theorie+, principe+, condition+, rapport+, difference+, cas, experience+, element+, systeme+, forme+, realite+, social+, probleme+, sens, loi+, valeur+, phenomene+, action+, physique+, politique+, developpement+, terme+, agit, objet+, analyse+, nature+, hypothese+, concept+, divers+, representation+, question+, economique+, matiere+, essentiel+, methode+, unite+, reel+, science+, sujet+, philosophie+, domaine+, fonction+, repetition+, physicien+, activite+, puissance+, evolution+, commun+, exemple+, mathematique+, opposition+, langage+, different+, concret+, national+, consequence+, francais+, resultat+, conception+, moderne+, logique+, generale+,

Tableau 2. Traces des 4 mondes lexicaux stabilisés servant d'étalonnage pour la version 5 du logiciel

Commentaire (cf. Tab. 2). Est-ce un hasard si la trace du monde lexical lié à l'imaginaire est introduite par le mot « ombre » ? Sans doute, mais l'évocation du mythe de la caverne de Platon augure bien de ce que l'on cherche à signifier ici par « imaginaire », c'est-à-dire, de l'impossible certitude de l'existence d'un monde qui ne serait accessible qu'à travers les sens, et qui est pourtant antérieur à notre propre naissance ? Nous rejoignons là Peirce pour qui le monde de la priméité est celui de la simple possibilité d'être, et non celui de l'existence, accessible en actes. Dans les œuvres plus spécifiques de ce premier pôle dominent celles de Victor Hugo. Le genre discursif le plus prégnant est celui de la poésie.

Les deux mondes lexicaux suivants (classes 2 et 3) sont l'expression dans le vocabulaire des deux bords d'un présent caché et pourtant là à chaque instant, innommable donc comme catégorie ou registre : le réel[23]. Il se montre justement à travers ce vacillement constant entre passions et actions, affects et engagements. Les mondes lexicaux associés sont marqués par les oeuvres romanesques retenues (Balzac, Céline, côté « Réel-socius » ; Laclos, Rousseau côté « Réel-affect »).

Enfin, le monde lexical soutenant la classe 4, différencié dès le premier pas de la classification, évoque un vocabulaire de concepts (philosophique, scientifique, du moins pour les termes les plus saturés). Il est remarquable de trouver dans les œuvres les plus caractéristiques de ce pôle, les discours du général de Gaulle, au côté des œuvres de Deleuze, Duheim, Bergson et Saussure…

Cette analyse a ainsi permis de classer 8 865 *formes réduites* en 4 registres : A, B, C, ou D selon leur appartenance privilégiée (au sens du $\chi 2$ à 1dl) à l'un des 4 mondes lexicaux de ce corpus d'étalonnage[24]. En voici une statistique, réduite aux seules formes classées comme noms, adjectifs ou verbes (soit sur 8 643 formes réduites), les noms et adjectifs n'étant pas différenciés entre eux[25] :

	A (imaginaire)	**B (réel-affect)**	**C (réel-socius)**	**D (symbolique)**
Noms & adjectifs	**2 048** ($\chi 2 = 35$)	621	902	**1 843** ($\chi 2=10$)
Verbes	1 018	**596** ($\chi 2=81$)	**627** ($\chi 2=10$)	988
Marge	***3 066***	***1 217***	***1 529***	***2 831***

Tableau 3. Distributions des formes entre leur appartenance à un monde lexical et leur catégorie grammaticale

Commentaires : Le monde lexical A différencie nettement plus de noms ou d'adjectifs, que les mondes lexicaux B et C, plus associés à des verbes. Ceci semble montrer que les verbes même lemmatisés se comportent davantage comme des déictiques, si l'on accepte l'idée que les mondes lexicaux B et C sont globalement plus en rapport avec cette dimension indexicale du sens (la monstration du réel par opposition à la démonstration du symbolique à l'aide d'une représentation et à la désignation de l'imaginaire par catégorisation). Mais ceci n'est qu'une ouverture pour une réflexion future…

[23] Dans notre conception, qui vient de Lacan et de Peirce, le Réel fait trou dans le tissu des habitudes, dans la continuité de l'Imaginaire, dans ce qui constitue notre monde hérité du passé. En ce sens le réel s'impose contre la croyance, contre l'évidence des sens. Il ne se laisse pas domestiquer par les formes de ce dont on a déjà l'expérience. Lorsque ce tissu de croyances cède, que le réel s'impose comme violence, le symbolique est ce par quoi une suture pourra être pratiquée au prix d'un changement d'habitudes, au prix de l'acceptation d'un nouvel état des choses, etc.

[24] Ces registres correspondant également à des types de discours, à des topiques, aussi, nous appelons les lettres, A, B, C et D, en tant qu'elles affectent les mots du dictionnaire à un de ces 4 registres, des *« clés topiques ».*

[25] Par exemple, 2048 formes réduites différentes du monde lexical « imaginaire » ont été classées comme nom ou verbe : Soit un $\chi 2$ à 1 degré de liberté de 35, qui indique que 2048 est nettement plus élevé qu'attendu. On remarque d'autre part que la distribution des formes entre I, R, et S est assez bien équilibrée (environ 3000).

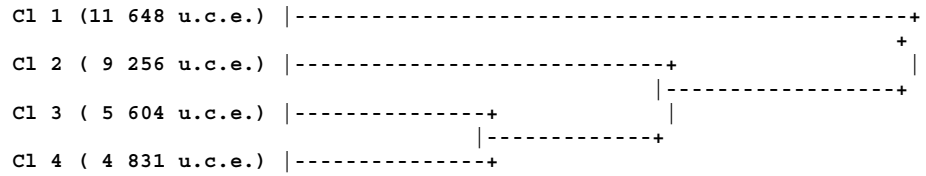
3.3 L'analyse d'un dictionnaire Larousse

Présentons maintenant rapidement l'exemple d'une utilisation des clés topiques[26] avec l'analyse d'un dictionnaire courant[27], analyse effectuée avec le plan standard d'*analyse simple sur les U.C.E.* de la version 5, sans préparation particulière du corpus[28], ni modification du plan, sinon d'imposer un minimum de 4 800 U.C.E. par classes, après un premier essai, afin de réduire le nombre de classes terminales à 4. En voici les principales caractéristiques :

Nombre de Segments de Texte Calibrés	87 236
Nombre de caractères	16 831 415
Nombre d'occurrences	2 531 105
Nombre de formes distinctes	87 626
Nombre de formes réduites communes avec l'analyse étalon	6 043[29]
Nombre de formes réduites analysées	2 987
Nombre de mots supplémentaires	6 441
Nombre d'occurrences par U.C.E.	60
Nombre d'unités de contexte élémentaires (U.C.E.)	34 193
Nombre de classes demandées	15
Nombre minimum d'U.C.E. par classe (demandé)	4 800
Nombre d'occurrences analysées	593 434
Poids du tableau effectivement analysé par la C.D.H.	533 764
Nombre d'U.C.E. classées par la C.D.H.	31 339 soit 91.65%
Nombre d'U.C.E. classées dans une « classe topique »	26 804
Nombre d'U.C.E. communes aux 2 classements	24 804

Les statistiques portent ensuite sur les seules 31 339 U.C.E. classées par la C.D.H.

L'arbre de la Classification Descendante Hiérarchique :



Remarque : Le découpage en articles est marqué par une ponctuation spéciale pour éviter un recouvrement pour ce qui concerne le découpage en segments de texte calibrés[30] : 87 236 S.T.C. sont distingués. Ils sont réunis ensuite dans 34 193 U.C.E.[31], utilisées pour les statistiques ; Ces U.C.E. peuvent donc recouvrir localement des définitions différentes, mais de bases lexicales souvent proches. La classification statistique (C.D.H.) a porté sur un tableau

[26] Nous appelons « clés topiques », les 4 lettres « A, B, C, D » affectées aux formes réduites en fonction de leur appartenance privilégiée à l'un des quatre mondes lexicaux de l'analyse-étalon.

[27] Petit Larousse, édition 1997.

[28] Les différentes définitions étaient déjà séparées par deux sauts de ligne successifs et nous en avons tenu compte pour séparer les « segments de textes calibrés ».

[29] 5732 formes réduites sont marquées par une clé topique (A, B, C ou D), une fois ôtés les mots outils. C'est à partir de la distribution de ces mots dans les U.C.E. que sont calculées les classes topiques. Ces classes sont ensuite comparées aux 4 classes obtenues ici avec la C.D.H. (cf. Tab. 4)

[30] Ce n'est pas le cas pour les U.C.E.

[31] Avec *ALCESTE*, le corpus est d'abord découpé en S.T.C. (d'au plus 240 caractères), en fonction de la ponctuation. Les S.T.C. sont ensuite concaténés dans des U.C.E. constituant les unités statistiques de base pour l'analyse. Chaque U.C.E. « mesure » ici environ 60 occurrences.

simple comprenant, en lignes, ces U.C.E., et en colonnes, les 2 987 « mots pleins » analysés (99% de zéros).

Sont calculées ensuite les classes topiques à partir du dictionnaire étalonné, ce qui permet de distribuer 26 804 U.C.E. (sur les 34 193) dans 4 classes topiques[32], dont 24 804 U.C.E. se trouvent être également classées par la C.D.H. Le $\chi 2$ d'association[33] des classes topiques avec les classes de la C.D.H. est très élevé (cf. Tab. 4).

La hiérarchie des classes ne correspond cependant pas avec celles de l'analyse précédente, la classe la mieux discriminée ici étant celle associée à la classe topique C (Réel-socius). Elle correspond en fait à la partie du dictionnaire, *relative aux noms propres,* dont le style est plus narratif.

La seconde classe discriminée est liée à la classe topique A (Imaginaire). Elle réunit le vocabulaire des activités usuelles, matérielles, y compris le vocabulaire relatif aux sciences de la nature. Le style des définitions est la description.

Ensuite se séparent les classes 3 et 4, différenciant les classes topiques B (Réel-affect) et D (Symbolique). Le vocabulaire des affects et des termes abstraits nécessite des formes de définitions plus diversifiées (exemples, discussions). Notons la présence des valeurs morales dans la classe 3 rappelant que valeurs et jugements ne sont pas indépendants des affects.

Nous n'approfondirons pas plus cette analyse dans le cadre de cette présentation. Ajoutons cependant que Margareta Kastberg (2008) présente dans ces journées une analyse sur la partie « française » d'un dictionnaire bilingue français-suédois. Une analyse standard effectuée avec la même version du logiciel[34] a permis (au premier essai) de mettre en évidence 3 classes stabilisées, ces trois classes étant très nettement liées aux classes topiques A, B et D. La classe liée la classe topique C manque. Dans notre analyse, cette classe est associée à la partie du dictionnaire recouvrant les articles liés aux noms propres, partie justement absente de ce dernier ! Cela dit, une autre partie aurait pu s'y substituer… ce n'a pas été le cas, suggérant un rôle non négligeable des noms propres (en tant qu'entrées du dictionnaire[35]) pour exprimer la singularité de ce registre (marqué par la clé topique C dans le dictionnaire étalonné).

Ce champ de recherches est à peine exploré. Il est très largement ouvert, et concerne aussi l'approche des genres littéraires comme les recherches d'E. Brunet et de son école le suggèrent[36]. Nous terminerons là-dessus. Merci de votre attention.

---

[32] Le calcul consiste grossièrement à distribuer les occurrences de chaque U.C.E. en fonction des 4 clés topiques et à affecter l'U.C.E. dans la classe topique correspondant à la clé topique la mieux représentée (au sens du $\chi 2$), pour les occurrences associables à un mot classé dans le dictionnaire étalonné.

[33] Rappelons que nous appelons $\chi 2$ d'association la valeur d'un $\chi 2$ sur un tableau de contingence 2x2, affectée du signe du produit des 2 termes diagonaux après soustraction du produit des 2 termes antidiagonaux.

[34] Avec le plan standard (avec double analyse). Les trois classes présentées ont été obtenues dès le premier essai. Ce plan n'a pas été utilisé pour l'analyse du Larousse du fait, entre autres, de la taille plus importante du corpus. L'analyse simple est préférable alors si l'on désire faire l'analyse sur les unités textuelles les plus réduites possibles.

[35] Les mots en majuscules ou commençant par une majuscule ne sont pas analysés ici. Ce sont évidemment les discours introduits par les noms propres qui sont liés à la classe topique C.

[36] La méthodologie d'E. Brunet (avec le logiciel Hyperbase) se rapprochant de la nôtre quand il fait l'analyse des correspondances de tableaux croisant lexique par pages, tableaux construits à partir de corpus d'œuvres littéraires (la page jouant alors le rôle d'une unité de contexte).

Classe 1 (11 648 U.C.E.) { CT_C (Réel-socius) : 2 709 U.C.E. sur 5 009 ; $\chi 2$ = 730 }	politique+, roi+, guerre+, musee+, pays, canton+, president+, nord, royaume+, ministre+, britannique+, americain+, roman+, province+, sud, ville+, allemand+, auteur+, ile+, fit, dynastie+, chef+, independance+, republique+, devient, capital+, comte+, romain+, departement+, devint, empereur+, duc, saint+, traite+, gouvernement+, ouest, elu+, proclam+er, principal+, peuple+, national+, chretien+, gothique+, chef-lieu, russe+, population+, frere+, general+, eglise+, elevage+
Classe 2 (9 256 U.C.E.) { CT_A (Imaginaire) : 4 991 U.C.E. sur 7 770 ; $\chi 2$ = 5 977 }	utilis+er, servant, bois, eau, peau+, liquide+, plante+, metal+, fleur+, petit+, animal+, surface+, metallique+, blanc+, feuille+, tissu+, espece+, fruit+, arbre+, vegetal+, acide+, couleur+, termin+er, papier+, contient, tige+, jaune+, long+, corps, sol+, verre+, oeuf+, pierre+, singulier+, gaz, taille+, creu+x, etoffe+, insecte+, graine+, matiere+, semblable+, lame+, cultiv+er, color+er, fer, parlant, pointe+, air, oiseau+, grain+, plat+, piece+, cavite+, rouge+, os, pate, epais+, sang+, mince+, patte+
Classe 3 (5 604 U.C.E.) { CT_B (Réel-affect) : 2 233 U.C.E. sur 4 235 ; $\chi 2$ = 4 048 }	etre, adverbe+, faire, facon+, avoir, mal, bien, indirect+, fait+, sentiment+, chose+, etat+, acte+, maniere+, moral+, agir, bonne+, mauvais+, pouvoir., situation+, mettre, manque+, prendre, dire+, affaire+, parole+, peine+, locution+, parler, obligation+, cause+, preuve+, danger+, excessi+f, attention, juge+, perdre, droit+, plaisir+, rendre, jugement+, justice+, aller, trouver, decision+, venir, evenement+, donner, sante, manifeste, penible+, difficile+, consequence+, action+, manifester, enfant+, debiteur+
Classe 4 (4 831 U.C.E.) { CT_D (Symbolique) : 2 786 U.C.E. sur 7 759 ; $\chi 2$ = 3 321 }	ensemble+, element+, systeme+, operation+, image+, unite+, donnee+, phenomene+, information+, valeur+, personnes, nombre+, mesure+, permettre., programme+, fonction+, television+, determine, entreprise+, service+, grandeur+, different+, relati+f, activite+, etude+, enregistr+er, appareil+, etudie, emission+, distance+, methode+, etres, variable+, radio+, espace+, chimique+, propriete+, reel+, technique+, rapport+, document+, signal+, numerique+, structure+, science+, automatique+, produit+

Tableau 4 : Traces des quatre principaux mondes lexicaux du dictionnaire Larousse[37]

Références

Achard P. (1991). Une approche discursive des questionnaires : l'exemple d'une enquête pendant la guerre d'Algérie. *Langage & Société,* 1991, n°55 : 5-40.

Balat M. (2000). *Des fondements sémiotiques de la psychanalyse.* L'Harmattan.

Benzécri J. P. & Coll. (1973). *L'analyse des Données.* Dunod.

Brunet E. (2006). Navigation dans les rafales. In *JADT 2006*, Besançon. (http://www.cavi.univ-paris3.fr/lexicometrica/jadt/JADT2006-PLENIERE/JADT2006_EB.pdf).

Everaert-Desmedt N. (1990). *Le processus interprétatif.* Ed. Pierre Mardaga, Bruxelles.

Kastberg M., Reinert M. (2008). Le discours dictionnairique : analyse systématique des structures sémantiques. In *JADT 2008*, Lyon.

Peirce Ch. S. (v.f. 1978, Traduit et commenté par G. Deledalle). *Écrits sur le signe.* Éditions du Seuil.

Reinert M. (2007). Postures énonciatives et mondes lexicaux stabilisés en analyse statistique de discours. *Langage & Société*, n° 121-122 : 189-202.

Reinert M. (1997). Les 'Mondes lexicaux' des six numéros de la revue 'Le Surréalisme au Service de la Révolution'. *Mélusine,* XVI : 270-302. Editions L'Age d'Homme, Lausanne.

Reinert M. (2001). Processus catégorique et co-construction des sujets et des mondes à travers l'analyse statistique de différents corpus. In *Linguistique et psychanalyse* [Actes du colloque de Cerisy-la-salle, sept. 1998 organisé par M. Arrivé & Cl. Normand]. Editions In Press, 385-397.

[37] Les listes de mots analysés du tableau sont ordonnées par $\chi 2$ d'association décroissant.