Rêves déçus et espoirs raisonnables à propos d'un logiciel de classement automatique en philosophie

Alain Lelu, Benoît Hufschmitt

Rêves déçus et espoirs raisonnable à propos d'un logiciel de classement automatique en philosophie

Benoit HUFSCHMITT, LASELDI, maître de conférence à l'Université de Franche-Comté.
Alain LELU, LASELDI, professeur à l'Université de Franche-Comté

Ces notes sont sans prétention théorique, elles ne visent pas non plus à faire la publicité ou la critique d'un logiciel spécifique d'analyse textuel. Leur but est simplement de tenter de dégager ce qui peut être espéré et ce à quoi il faut renoncer avec un tel logiciel, dont la spécificité est de manipuler des quantités importantes d'unités textuelles déclarées, laissant les formes plus classiques d'analyse des données manipuler des unités moins nombreuses.

Un quiproquo préalable.

Travaillant depuis quelques années sur une synthèse des apports que l'informatique peut apporter à l'étude des textes philosophiques, entre autre au niveau documentaire, je <BH> recherchais des applications lexicologiques qui puissent me permettre d'intégrer divers usages importants relativement à mon projet :

Un outil de recherche lexicale évolué

Je désirais d'abord trouver un analyseur sémantico-syntaxique qui produise une normalisation du vocabulaire des textes au niveau des formes " dictionnaires " afin de permettre des recherches lexicales plus fines que le plein texte, recherche de thèmes à l'intérieur d'un corpus d'auteur (ce sera ici : l'édition Adam-Tannery des *œuvres complètes* de Descartes, ensemble à peu près numérisé par un éditeur américain de cdroms d'œuvres philosophiques : *Past Master*).

Du côté d'un chercheur en histoire de la philosophie, l'intérêt d'une telle approche, surtout si elle autorise l'utilisation de formules booléennes, est immédiat : extension des recherches plein texte avec moins de bruits (homographie réduite) et moins de silences (formes graphiques ignorées moins nombreuses).

Un outil de recherche conceptuelle

J'espérais aussi acquérir une (ou plutôt des) classifications automatiques du vocabulaire d'un corpus, via les outils d'analyse des données, afin de permettre que soient automatiquement proposées des classes de vocabulaire qui pourraient recouvrir au moins partiellement (ce qu'il faut dire avec beaucoup de prudence) le vocabulaire des champs conceptuels d'une doctrine philosophique. J'ajoute immédiatement que je reste, à ce jour très sceptique sur ce genre de résultats, du moins en l'absence de stratégies d'approche sérieuses.

Pratiquement, cela consisterait à pouvoir disposer d'une liste de liste de vocabulaire unifié sous un concept, nommé par un élément de vocabulaire.

Je pensais pouvoir en faire, plus précisément, trois usages distincts :

- Dans le prolongement de ce qui précède : recherche de passages précis à l'intérieur du corpus concernant le concept choisi (ou une composition booléenne).
- Dans une perspective de validation de propositions faites par les commentateurs, voire d'évaluation des commentateurs ; la réciproque s'impose d'abord, direz-vous, certes, mais cela ne relève pas de ce cadre.
- Comme auxiliaire d'invention ou de découverte : propositions d'unités conceptuelles à considérer et analyser humainement ensuite, afin de les prendre en considération pour un travail de recherche.

Un outil d'organisation conceptuelle

Dans la continuité de ce travail de classification automatique, je pensais pouvoir disposer de classifications des concepts et, indirectement, du vocabulaire lui-même, permettant d'affiner la recherche de passages thématiquement fixés, soit par des concepts, soit par le vocabulaire, en maîtrisant la sélection du vocabulaire proche : par mise à disposition d'un tableau d'analyse factorielle qui rapproche les éléments de cette liste de liste selon une proximité bi-dimensionnelle, voire par un graphe de proximités, sans espérer atteindre une structure mieux ordonnée (arborescence ou treillis).

Je n'osais penser que cette organisation puisse être déterminée plus finement qu'une simple proximité, afin de déterminer des types de relations et donc une ébauche de caractérisation terminologique.

Un outil de classement des textes eux-mêmes

Enfin, je cherchais le moyen de produire des classes de documents proches dans leurs objets d'étude, à partir de leur proximité de vocabulaire, le passage par des descripteurs (documentaires ou conceptuels) pouvant être omis. Des cartes d'analyse factorielle dynamiques devraient en ce cas pouvoir être construites par paramétrage de focalisation selon certains textes, voire selon du vocabulaire spécifique. Cette démarche, nonobstant la qualité des résultats, est très proche de ce que propose le plus couramment un commentateur : des références à d'autres textes du corpus, à la différence évidemment que la machine n'a effectué ce rapprochement que par analyse du vocabulaire, alors que le commentateur a mis en œuvre une composition subtile de raisons où le vocabulaire n'est pas absent, mais non pas dominant.

Deux types de préoccupations se mêlaient donc : d'ordre théorique lorsqu'il s'agit de délimiter des concepts par le vocabulaire, voire de cerner leurs relations ; d'ordre documentaire dans les autres cas, encore que, par exemple, organiser les textes les uns par rapport aux autres relève aussi de préoccupations théoriques.

Accès au logiciel Neuronav

C'est dans cet esprit que je pris contact avec le professeur Lelu, concepteur du logiciel Neuronav commercialisé par la société *Diatopie*, " logiciel de *Text Mining* consacré à l'analyse statistique, à la classification et au parcours de vastes ensembles de documents textuels aux formats usuels (Pdf, Office, Txt, Html) ", à destination du monde professionnel avant tout.

Ce logiciel est présenté ainsi en page d'accueil du site internet [1] qui le concerne :

Neuronav vous aide à :

- *Prendre connaissance rapidement et globalement du contenu des documents qu'il serait inenvisageable de lire un par un.*
- *Mettre au point une terminologie propre à votre domaine de connaissance.*
- *Classer les documents.*
- *Observer la façon dont évoluent les contenus des documents d'une même source au cours du temps.*
- *Comparer les contenus provenant de plusieurs sources différentes.*

M. Lelu, cependant, s'intéressait plutôt aux usages qui pourraient en être fait dans le monde universitaire, et était à la recherche d'expérience documentaires théorisées et d'évaluations plus complexes que les satisfactions d'usage dans la gestion d'archives commerciales ou administratives. C'est ainsi que, apprenant l'orientation de mon propre travail, il me proposa de travailler le corpus philosophique qui me sert depuis quelques années de terrain d'expérience : l'œuvre de Descartes.

Fonctionnement [2]

Ce logiciel étudie un corpus de textes (mono ou pluri-fichiers, ici tout le corpus a été groupé en un fichier) dont il demande un découpage en de nombreuses unités sémantiques larges (lettres, articles de journaux, rapports, poèmes …) représentant les passages qui seront comparés [3].

Il construit ensuite le vocabulaire (normalisé mots de dictionnaire) du corpus, éliminant au passage les mots jugés fonctionnels déposés dans un anti-dictionnaire (dynamiquement modifiable à tout moment du travail). Cette liste inclut ce qui a été repéré comme " mots composés " à partir d'un patron syntaxique et des listes d'exceptions. Cet analyseur peut être aisément amendé ponctuellement.

L'utilisateur peut alors forcer la sélection de ce vocabulaire, en éliminant généralement ce qui risque de parasiter les rapprochements (par exemple les mots " chapitre ", " article ", les titres, les en-têtes et formules de politesses des lettres…, les erreurs manifestes d'analyse (un panier de mots aide à ces opérations). Nous avons pris le parti, pour notre part, via une manipulation possible, de ne retenir que le vocabulaire propre à un petit texte : le *Discours de la Méthode*, puisque nous voulions des résultats focalisés par ce texte.

Ce nettoyage opéré, l'utilisateur peut choisir de ne considérer qu'une partie de l'ensemble de ce vocabulaire partitionné en substantifs, adjectifs, verbes et mots composés. Divers essais nous ont conduit à analyser séparément les substantifs, adjectifs et formes composées d'une part, formes verbales d'autre part. On choisit aussi un seuil de fréquence minimale, en général 2 (afin d'éliminer les hapax et les nombreuses formes composées non répétées).

---

[1] http://www.diatopie.com/ProdNeuroNav.htm.

[2] Description succincte in http://www.diatopie.com/NNdescription.htm.

[3] Pour l'étude d'un moindre nombre d'unités, l'analyse factorielle classique semble à privilégier.

Puis l'utilisateur détermine le nombre de nœuds thématiques qu'il désire voir construire, et peut alors lancer l'analyse de la table de fréquence des mots et la préparation des résultats sous forme d'une cartographie où les liaisons entre les nœuds (thèmes) sont représentées par des traits plus ou moins épais.

Résultats, aspect formel

Les résultats du travail sont divers et diversement paramétrables.

1- Le plus simple est la liste du vocabulaire, sur laquelle peut être opérée une sélection simple ou booléenne qui fournit en regard les passages qui la possèdent. mais ce n'est en fait que le matériau pour le véritable traitement (et un moyen empirique de contrôle).

2- Suite au travail dit de cartographie, l'analyse a déterminé les nœuds (thèmes) du corpus dans le nombre demandé, chacun représentatif d'un ensemble d'unités textuelles, illustré par une liste de mots caractérisants. ce qui donne accès à

   a. La liste des thèmes avec leur vocabulaire associé,

   b. La liste des unités textuelles liées à une sélection quelconque de mots ou de thèmes (ou leur composition booléenne).

   c. Une carte présentant les nœuds, les arcs apparaissant à demande selon l'importance calculée des liaisons. Un option permet de voir sur chaque nœud les mots les plus significatifs du thème.

3- Diverses opérations peuvent alors être faites sur ces résultats. Une sélection dans la liste des mots ou de celle des thèmes fournit un éclairage de la carte selon cette sélection. Dans le cas des thèmes, les fragments de texte particulièrement concernés sont mis en avant.

Difficultés de préparation des données

La mise en place des données révéla quelques difficultés, finalement rédhibitoires envers certains de nos espoirs.

Choix du corpus.

il semble simple de décider de traiter d'un corpus tel le corpus cartésien, car une édition princeps existe (édition Adam-Tannery du début du XXième siècle, révisée jusqu'en 1964), dont nous suivons, sans discussion, les décisions éditoriales. Que les documents informatisés retenus (le texte de l'édition fournie par Past Master) s'autorisent quelques libertés est un problème factuel.

Pourtant diverses difficultés se présentent relativement au fait que l'étude porte strictement sur le vocabulaire et que les traitements sont statistiques ; des décisions sont à prendre, plus ou moins difficiles à justifier, parfois lourdes de conséquences dans les calculs :

- La *Correspondance* inclut les lettres des correspondants de Descartes et les *Objections aux Méditations* précèdent évidemment les *Réponses* de Descartes. Les éliminer semble aller de soi ; mais on objectera que le vocabulaire est fortement déterminé par celui de Descartes et que les problèmes ou concepts en jeu sont ceux de Descartes. On pourra certes alors répondre que l'ensemble des textes de commentaires est alors pertinent, ce qui, outre la difficulté technique, noierait évidemment le discours strictement cartésien. C'est pourquoi, nous les avons enlevées.[4]
- D'autres auteurs interviennent souvent au niveau des préfaces, faut-il les éliminer de la même manière ? Nous les avons gardées en raison, outre le peu de volume qu'elles représentent, du fait que Descartes les a lui-même voulues (lettres servant de préfaces aux *Passions* par exemple).
- Il y a, dans la correspondance encore, de multiples sources de bruit, quand le texte sort évidemment du contexte de recherche philosophique. Si l'on peut conserver les cas, discutables déjà, des informations événementielles privées ou publiques, il est difficile de faire abstraction des en-têtes et formules de politesse qui risquent de peser de manière inopportune sur les calculs. Il semble donc préférable d'éliminer aussi ces éléments, du moins ceux qui se présentent clairement et peuvent être réduits par un traitement général. La réponse à cette difficulté consistant à éliminer du vocabulaire pertinent les mots du genre " cher " " révérent " " veuillez " " monsieur "… ne semble guère amputer le lexique pertinent et peut être aussi envisagée.
- Une partie importante du corpus est écrite en latin, une toute petite partie en hollandais ou anglais. Il est exclu, évidemment de les retenir, mais quelle pertinence y a-t-il à les remplacer ? Divers cas de figures se présentent.
  - Traduction française de Descartes, ou du moins revue et approuvée par Descartes. C'est le cas de *Méditations*, des *Principes*, des *Réponses* hors les cinquièmes et septièmes.

[4] Dans le même ordre d'idée, que faire des contenus de citations ? rares chez Descartes, elles ont été conservées.

- Texte substitutif à l'original latin écrit en français par Descartes, c'est le cas des *Cinquièmes Réponses* fortement amputées pour l'occasion.
- Traductions françaises non contrôlées par Descartes. Lesquelles prendre, quelles incertitudes engendrent-elles au niveau du vocabulaire ? S'il y a des cas où une édition " quasi-officielle " existe : la *Correspondance* éditée par Adam et Milhaud par exemple, il en est d'autres où aucune traduction ne s'impose et où les variations de vocabulaire sont importantes, engageant sans doute des choix philosophiques (*Regulae*).
- Reste enfin deux cas très particuliers :
  - Celui de la *Dissertatio*, traduction tardive en latin du *Discours de la Méthode*, contrôlée par Descartes. Des modifications du texte français, sont-elles importantes ? ont été faites. Faut-il introduire dans le corpus une traduction française de la traduction latine ?
  - Celui de *La Recherche de la Vérité* dont nous sont parvenus le début en français et la fin en latin. En outre, ce texte étant un dialogue, il pourrait être proposé que le seul discours du représentant de Descartes soit retenu.

Face à ces diverses difficultés, nous avons pris le parti de limiter le corpus aux textes écrits en français dans l'édition Adam-Tannery, amputant en conséquence le corpus réel de nombreuses lettres (dont l'ouvrage imposant que constitue à lui seul l'*Epistola ad Voetium*), les *Vae* et *VIIae Objectiones*, la *Dissertatio* et surtout les *Regulae ad Directionem Ingenii.*

Validité des réductions des mots aux formes normales (de dictionnaires)

Un premier niveau de travail sur le lexique d'un corpus prend comme unités les mots, en traitant au mieux les mots composés et les mots fonctionnels considérés comme de peu d'intérêt sémantique. Les difficultés à ce niveau sont bien connues (homographes non discernables, détermination des mots fonctionnels, mots composés, termes anaphoriques) et il est inutile de s'y appesantir. Postuler comme unité lexicale toute suite de mots qui se répète exactement est judicieux pour calculer des proximités de documents, mais comment en tester la valeur comme unités conceptuelles ?

Par ailleurs, on pourrait espérer une réduction convenable des formes lexicales par analyses sémantico-syntaxiques alternées, mais l'analyseur, assez rudimentaire, se contente de mettre en rapport des formes et des étiquette, réglant normativement les cas d'homographie, sans proposer de détermination manuelle des cas. Il y a, en outre, production de formes inattendues, dues en particulier à une préférence pour les formes verbales en cas d'incertitude (par exemple terre est systématiquement renvoyé au verbe terrer, muscle à muscler etc.. Ce n'est qu'un point de moindre importance, un compromis acceptable en rigueur et économie d'interventions humaines, dans le cadre d'une analyse statistique sur un gros corpus, mais c'est sans doute très handicapant au niveau de nos préoccupations lexicologiques et conceptuelles.

Réduction du lexique de travail

Cette question de la sélection du vocabulaire pertinent a une portée générale et une portée spécifique au projet envisagé.

- Il semble nécessaire de disqualifier le vocabulaire fonctionnel (encore que l'on puisse le juger à peu près uniformément réparti, sauf en des cas sémantiquement particularisés, donc dignes d'intérêt), mais il n'est pas toujours clair de le choisir, à cause de la variation des usages d'un même mot (sans parler de l'homographie résiduelle). Dans notre exemple, le verbe *être* a le plus souvent valeur de copule (mot fonctionnel), sauf dans son emploi d'affirmation de l'existence, sans parler du substantif *être*.
- Mais il apparaît vite que toute une partie du vocabulaire (lexical) est susceptible de perturber l'analyse, le vocabulaire méta-linguistique (dans le sens de la fonction méta-linguistique du langage chez Jakobson), celui à fonction phatique (dont les en-tête, formules de politesse, titres et sous-titres) ou encore (pour un corpus orienté vers la connaissance) celui à fonction poétique, voire à fonction émotive. Finalement au moins, trois types de décisions sont à prendre :
  - Choix d'une typologie fonctionnelle du discours,
  - Sélection de ce qui est à conserver dans cette typologie,
  - Sélection du vocabulaire précis à éliminer alors.

  Si les deux premières relèvent d'une simple décision théorique, comment parvenir à un traitement automatisé de la dernière ?
- L'application présente offre le choix de retenir les substantifs, adjectifs, verbes, mots-composés ou l'une quelconque de leur union. Quelques tests rapides montrent que les résultats sont, en première analyse du moins, fortement différents (exemple 6 nœuds sont reliés linéairement avec les substantifs et

expressions composée, non reliés pour 2 et reliés de manière complexe pour 4 avec tout le vocabulaire). La présence de cette option est une bonne chose, mais sur quoi s'appuyer pour opérer une sélection ?

  - A posteriori après étude de diverses cartes pour évaluer ce qui est le plus prometteur de sens ? L'exemple précédent nous ferait alors choisir l'option substantifs, adjectifs et expressions composées. Le risque est alors grand de nous engager dans la facilité des travaux à résultats non réfutables !
  - A priori en fonction d'une théorie générale sémantique (dont nous ignorons tout !). Pour un travail à orientation conceptuelle, le bon sens aristotélicien se limiterait aux substantifs, cartésien aux substantifs et adjectifs, frégéen aux trois formes simples. Pour un travail documentaire, l'apport des expressions composées semble primordial.

- Si cette sélection dépend, ce qui semble guère discutable, des objectifs qui fondent la démarche, il nous manque les raisons précises qui permettent de déterminer précisément la liaison entre choix et objectifs. On soutiendra assez facilement que les objectifs documentaires de rapprochement de textes demandent tout le vocabulaire, mais qu'en est-il d'autres projets documentaires et surtout conceptuels.
- Un autre choix proposé est le seuil de répétition minimal de vocabulaire. Eliminer les hapax semble aller de soi [5], mais faut-il seuiller au dessus de 2 ? Aucune raison déterminante ne semble s'imposer (en rapport aux objectifs, au type retenu, à la taille du corpus) ?
- En outre, ces objectifs peuvent être aussi très spécifiques. Pour nous, par exemple, le travail est focalisé sur le *Discours de la méthode*. Nous avons choisi en conséquence de ne retenir que le vocabulaire de ce texte (hapax inclus par rapport à ce texte), sans aucune garantie ni même raison inclinante, sinon la sempiternelle impression d'atteindre des résultats intéressants [6].

Choix des thèmes

le nombre des thèmes qui seront déterminés par les algorithmes de calcul relève d'un choix de l'utilisateur. Une ébauche de choix est indiquée dans la présentation du logiciel : peu en cas d'objectifs synthétiques (il y a peu de classes !) mais cela ne mène pas loin. Il semble donc utile que l'utilisateur possède un pré-savoir sur les thèmes dominants du corpus (ou de la partie de corpus), mais cela reste problématique eu égard au fait que, encore cette connaissance serait-elle acquise, elle prête généralement à analyses ou synthèses diverses (amplifiant ou diminuant le nombre de thèmes retenus).

Pour le *Discours*, on peut d'abord juger que les six parties manifestent six thèmes ; mais la première partie se décompose clairement en trois thèmes distincts : " le bon sens ", les études, les voyages ; la cinquième présente quatre parties, certes liées : le monde céleste, inorganique, organique et la spécificité humaine (cette dernière à identifier peut-être au " bon sens ") etc. Il semble qu'il faille réfléchir tout d'abord à une liste générale de thèmes du corpus, ce qui est peut-être facile dans le cas de textes commerciaux ou administratifs, mais l'est moins pour un corpus théorique, le nombre étant de toute façon important. Un moyen peut-être est de réfléchir en termes disciplinaires, en y ajoutant quelques thèmes récepteurs des mots parasites. On aurait pour le Discours par exemple, en s'appuyant sur le célèbre arbre de la science de la *Lettre-Préface aux Principes* : métaphysique, physique (+ biologie), médecine, mécanique, morale, méthode, mathématiques, soit 8 + 2. De premiers essais avec ce nombre sont mitigés : 2 nœuds (liés) semblent représenter la biologie, 1 la physique et 1 les mathématiques (liés), 2 la métaphysique et 1 le " bon sens " (liés), 3 sont indéterminés (dont 1 lié au groupe précédent). Une carte avec douze nœuds donne des résultats de même qualité.

Point plus important, le travail de synthèse est fait à partir d'en ensemencement aléatoire de candidats termes, d'autres ensemencements fournissent d'autre réponses, comment gérer cette possibilité d'une diversité classificatoire qui n'est cependant pas arbitraire ? Il semble bien ici que l'on est contraint à une conception pragmatique de toute connaissance produite à ce niveau, méthodologiquement s'il n'y a d'ambitions que d'efficacité documentaire à rapprocher les textes ou de production de pistes de recherche, philosophiquement si l'on juge pouvoir atteindre ici une organisation conceptuelle du corpus.

Choix des unités de traitement

La préparation du corpus génère une dernière difficulté qui engage, nous semble-t-il, toute la valeur du traitement qui suivra. Le logiciel demande le découpage du corpus en unités réduites dont le sens n'est pas clairement fixé : unités sémantiques pressenties, unités éditoriales, ou nombre optimal de mots à solidariser a priori ! Ce second niveau d'unités sémantiques est nécessaire pour déterminer l'existence sinon la nature (ce qui serait l'idéal) des relations conceptuelles fortes, lesquelles donnent le contenu sémantique des concepts (faute de définitions caractérisantes de clarté et de distinction). L'unité minimale à laquelle on peut penser est la proposition élémentaire ; faute de pouvoir la déterminer automatiquement, on peut se retourner vers la phrase.

---

[5] Surtout pour un logiciel qui multiplie les formes composées.

[6] Ce qui fournit, à titre indicatif, hapax éliminés, 525 formes pour les substantifs, 1415 formes pour tout hors expressions composées, 1513 formes pour tout.

Cette réponse se révèle immédiatement insuffisante faute de pouvoir déterminer les marqueurs anaphoriques (ce qui est particulièrement perturbant à ce niveau) et en raison de la taille trop petite du vocabulaire rapproché (où les bruits liés à des formes rhétoriques et aux colocations sont importants). L'unité textuelle forte que représentent un chapitre d'ouvrage, un article, une lettre semble éviter le problème précédent. Mais il semble alors que l'ensemble soit trop vaste, regroupant en général diverses thématiques. C'est pourquoi, le paragraphe a été finalement retenu ici, d'autant plus que Descartes, inventeur du paragraphe d'après Vandendorpe, devait avoir une conscience aiguë de son bon usage.

Mais cela ne suffit pas encore car les paragraphes sont très différents de taille. Ce qui nous a conduit finalement à construire automatiquement des regroupements des plus petits paragraphes et des divisions des plus grands. Mais il a fallu alors, pour cela, déterminer des règles de choix pour ces modifications, l'une s'imposait : ne pas déborder, pour un regroupement, des limites d'une unité sémantique forte (lettre, chapitre), ne diviser qu'au niveau d'une changement de phrase ; pour le reste, la solution a été de choisir, lorsque c'était possible, les réunions ou coupures, par un premier calcul de proximité du vocabulaire.

Une autre solution a été envisagée, mais finalement délaissée : découper suivant une longueur relativement stable des textes, en se contentant des contraintes des fins de phrases et du non chevauchement d'unités sémantiques fortes.

Tout cela est finalement insatisfaisant. Il nous semble que la solution acceptable est :

- Soit de s'en tenir aux unités sémantiques fortes (lettres, articles, voire chapitres…) [7].
- Soit s'en tenir aux paragraphes, en coupant les trop longs et unifiant les plus courts.
- Soit délimiter le texte en fonction du plan décelé, ce qui contraint alors à un travail manuel de préparation considérable sur un gros corpus.

Dans tous les cas, pouvoir réduire le silence produit par les anaphoriques serait le bienvenu.

Nous avons finalement choisi la dernière solution en ce qui concerne le *Discours* lui-même, dont nous rappelons la place privilégiée puisqu'il est source du vocabulaire retenu. Et nous avons composé au mieux pour le reste du corpus afin parvenir à un découpage semi-automatisé : seuil minimum (1500 caractères) et seuil maximum (4000 caractères) de contenu textuel, limites impératives par les unités sémantiques fortes, limites selon les paragraphes divisés ou unifiés selon les nécessités en opérant les divisions à l'aide des marqueurs de phrases et les unifications en analysant succinctement le vocabulaire.

Rêves déçus

Il nous fallut donc, a priori, en rabattre sur nos espoirs initiaux.

- Un analyseur approximatif, sans être inutile, n'est pas la meilleure réponse pour des recherches lexicales.
- L'idée d'approcher le fond conceptuel du corpus cartésien a progressivement perdu toute crédibilité : à l'insuffisance de la réduction lexicale, s'ajoutent l'arbitraire dans les options de traitement, l'impossibilité à découper le corpus de manière satisfaisante, et surtout le relativisme des choix thématiques eux-mêmes, au cœur des algorithmes.
- Quant au rapprochement des textes, on ne pouvait en espérer qu'une incitation à en parcourir certain plutôt que d'autres, pour en comprendre un premier, sans assurance fine de la pertinence des rapprochements.

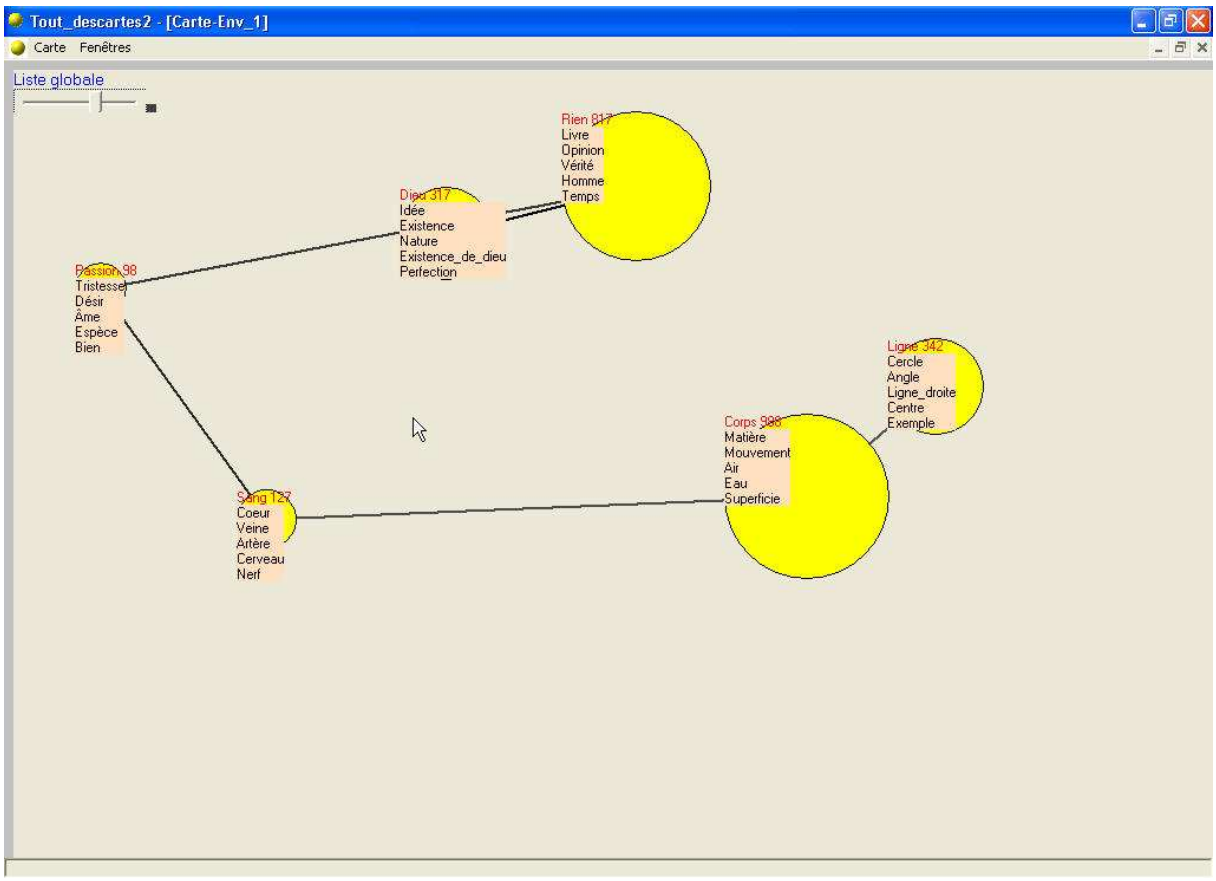
En bref, il ne semblait nous rester qu'à nous satisfaire d'une perspective documentaire ainsi que, cependant, d'une incitation à penser de nouvelles relations conceptuelles (arcs), voire de nouveaux concepts (nœuds).

Des espoirs raisonnables

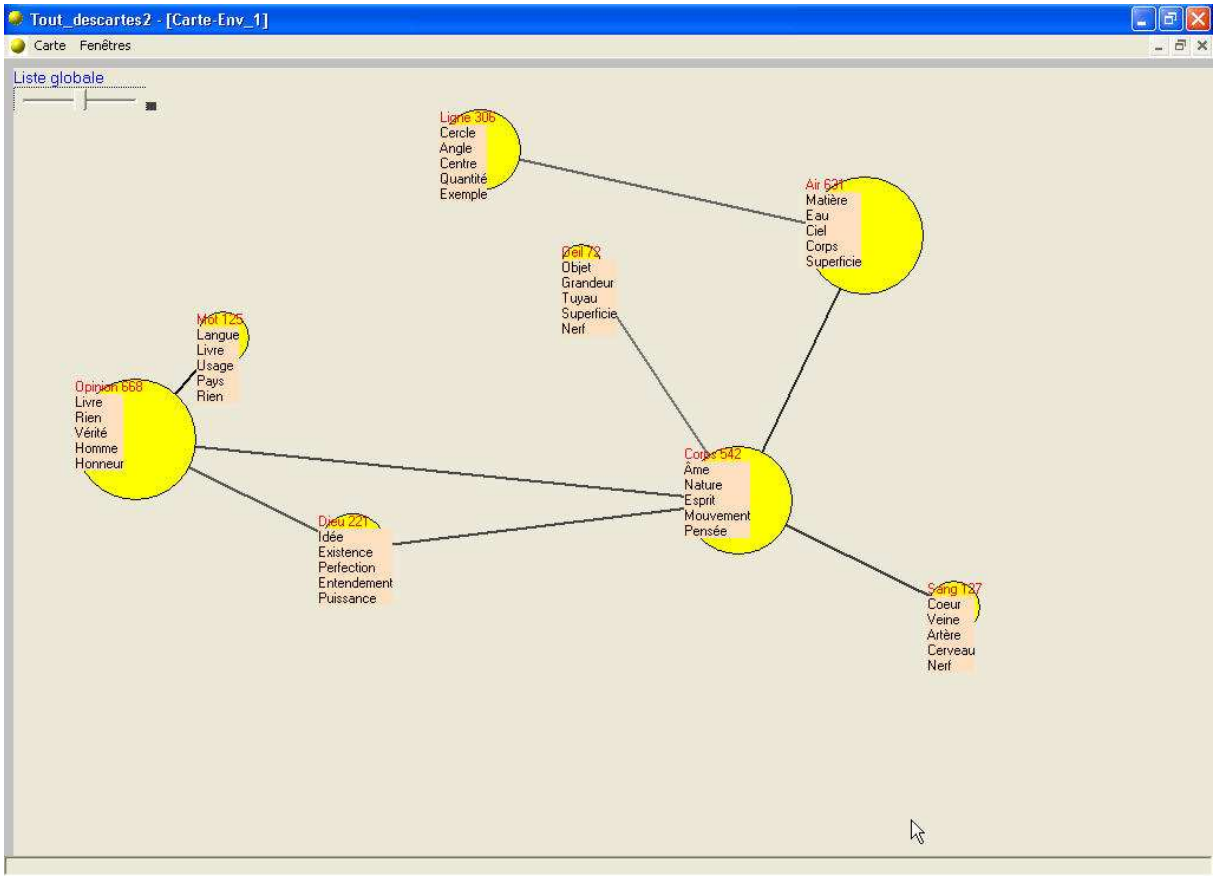
Les applications documentaires mises à part, que pouvons-nous donc faire de ces cartes et listes ? Le comportement spontané consiste à produire un grand nombre de résultats de l'application, par variation des options, des découpages, sélection du vocabulaire… et en extraire les plus intéressants, soit qu'ils confirment une idée connue sur le *Discours* ou le corpus, soit qu'ils proposent des pistes de lecture originales et relevant cependant du plausible ou du possible.

Du côté du connu, certains regroupements thématiques sont parlants à qui connaît un peu Descartes, de même que leurs liaisons. Par exemple la carte à 6 thèmes de substantifs et expressions composées où se déroulent les thèmes de Dieu, du savoir humain, des passions, de la biologie, de la physique jusqu'à la géométrie.

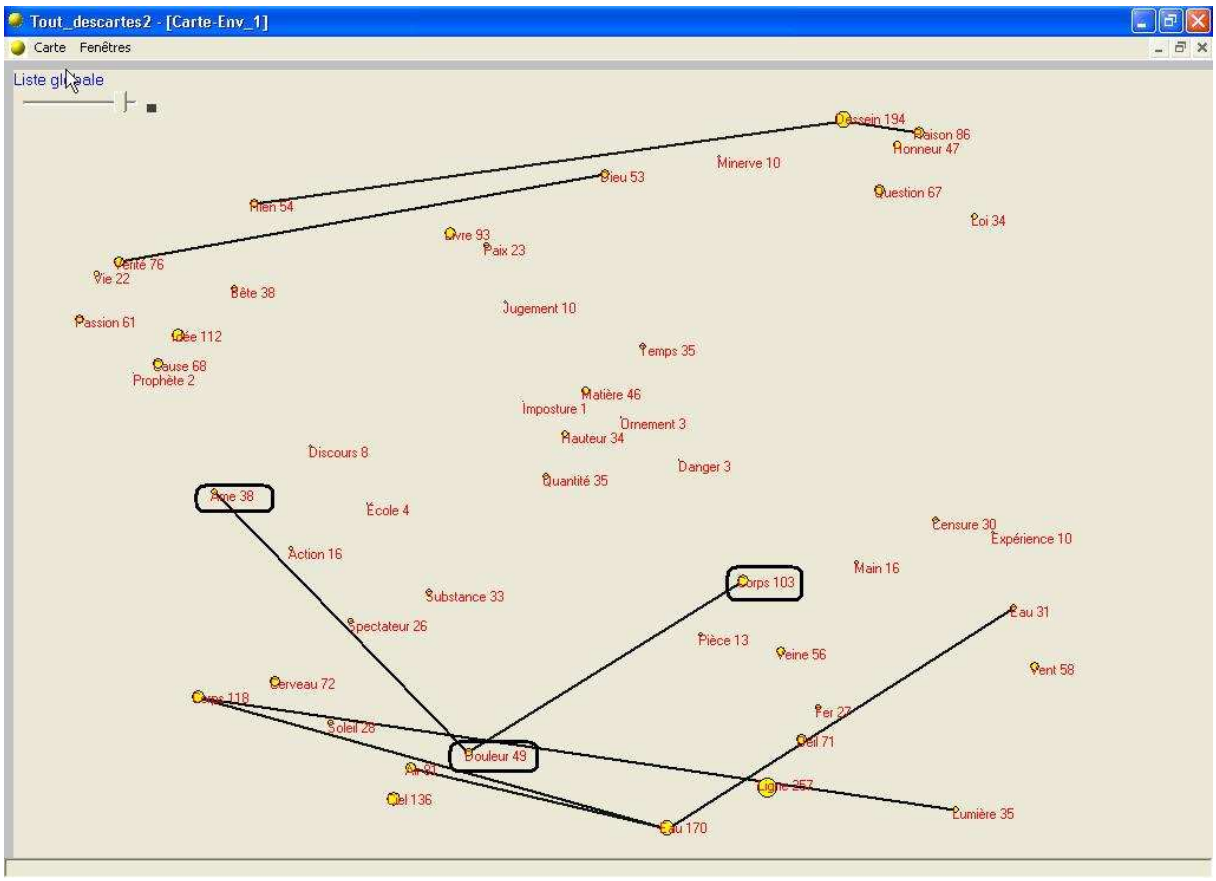
[7] Pour autant que leur nombre est important, sinon il vaut mieux utiliser l'analyse factorielle.



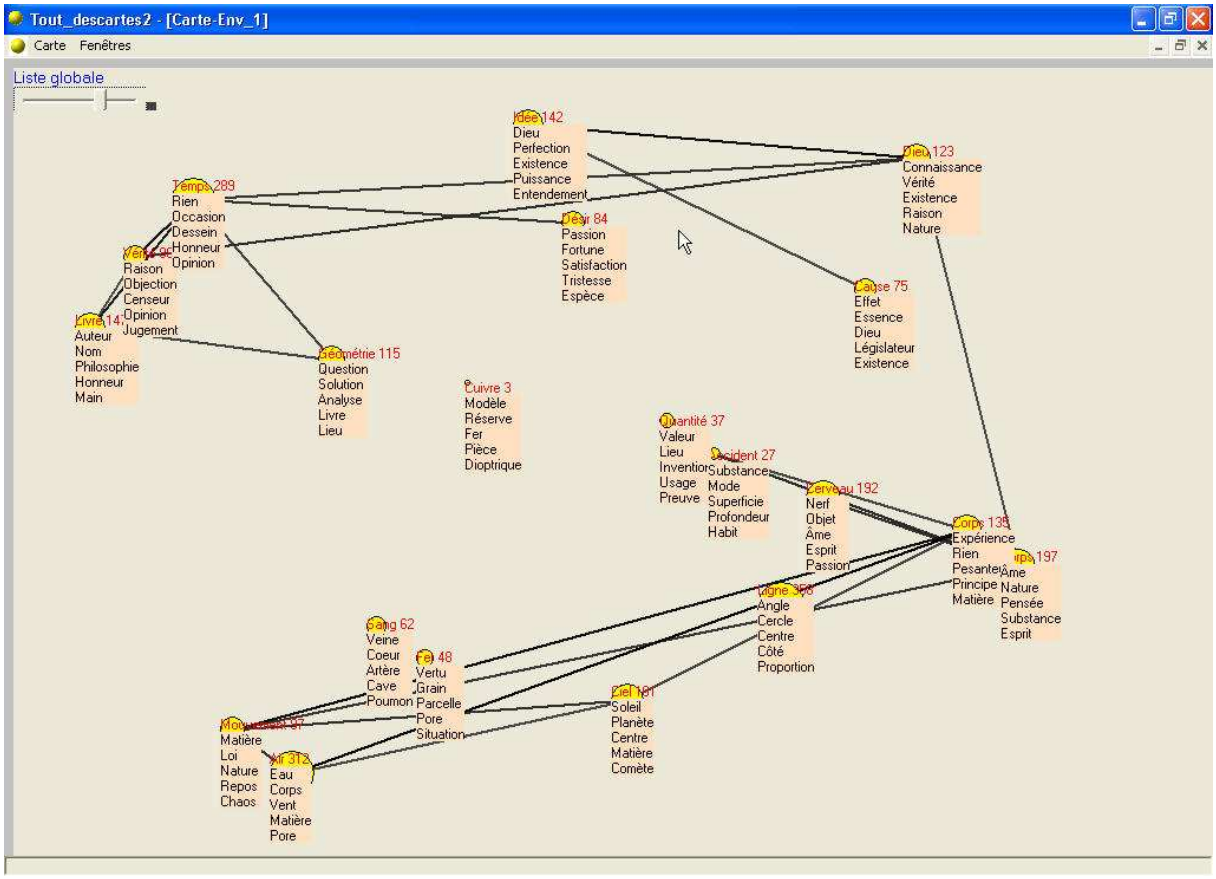
Mais pourquoi, ne pas préférer cette carte à 8 thèmes substantifs centrée sur l'union de substances qui irradie vers Dieu, l'humain, l'anatomie, la physique jusqu'à la géométrie.



Ce qui nous gène ici est que tout peut faire également sens (ou non sens). C'est encore plus net, évidemment du côté de l'inconnu : que penser, sur une carte de 50 thèmes, de ce que *douleur* est au centre d'une liaison forte entre *âme* et *corps* !



Ou encore, carte de 20nœuds, de l'opposition entre deux grands groupes thématiques qui n'ont de liaison qu'entre le thème Dieu-vérité et dualité des substances ?



Il se peut, certes, que cela aide à imaginer, à penser, mais qu'en tirer au niveau de la justification ou de l'argumentation ?

Outre la difficulté à envisager des possibilités de réfutation, il semble que l'on circule dans un cercle justificatif : l'interprétation des cartes valide l'application et l'application donne consistance à l'interprétation. On peut y répondre en développant la distinction entre *expliquer* et *justifier*, à l'instar de Descartes justement, dans la 6ième partie du *Discours de la Méthode*. Mais nous pensons plus intéressant de chercher à mettre en place des expériences ou l'invalidation a un sens. Nous suggérons celles-ci, dont nous pourrions fournir les résultats au mois de juin :

- Des études générales, par exemple :
    - Puisque l'on sait que la qualité de la lemmatisation, au niveau des mots normalisés et de leurs types, comme à celui des expressions composées est en soi peu satisfaisante, ne peut-on cependant pas tester ses effets sur la qualité des résultats produits. Une expérience peut être mise ici en place : lancer les mêmes recherches avant et après nettoyage fin du vocabulaire, les variations de résultats seules important.
    - Quelle est la stabilité des résultats. Y a-t-il influence de l'ordre des textes dans le travail de thématisation et les cartes ? La réponse théorique est oui en cas de traitement rapide, presque pas en cas de traitement fin. Que signifie ce " presque pas ", quels enseignements peut-on tirer des variations ? Pour ce faire, il suffit de présenter le même corpus différemment organisé (ou plusieurs fichiers proposés en un ordre différent).
- Des analyses dans le cadre du mode de présentation du corpus cartésien que nous proposons, construit comme indiqué ci-dessus (centré sur le *Discours de la Méthode*) :
    - Nous utiliserons le gros commentaire qu'Etienne Gilson a composé sur le *Discours* (étoffé, si possible, d'autres commentaires). Très précisément référé à l'ensemble du corpus cartésien, il doit donc pouvoir fournir les passages du corpus liés aux différents passages du Discours par une intelligence humaine, voire des générations d'intelligences puisque Gilson a fourni un travail très synthétique, lesquels passages seront confrontés à ceux que propose l'application.
    - Nous utiliserons le thésaurus que nous avons construit il y a quelques années déjà à partir d'un grand nombre d'index nominum d'œuvres de Descartes proposés par les divers commentateurs, pour le comparer aux thèmes proposés par l'application en faisant varier leur nombre.

  L'insuffisance des correspondances ne conduira évidemment pas à la contestation de la qualité du logiciel, mais à celle soit du logiciel, soit du travail de Gilson ou personnel. Sous cette dernière hypothèse, il s'avèrera alors intéressant de discuter de la valeur des résultats des travaux humains avec en arrière plan ceux de la machine.

Conclusion

Ce travail dont nous présentons ici le projet doit être compris comme le récapitulatif des difficultés à engager ce type de traitement statistique sur un corpus philosophique, à des fins de connaissance doctrinale ou d'organisation documentaire, ainsi que comme une réflexion d'ordre épistémologique sur l'utilisation de tels traitements à des fins théoriques. Sur ce dernier point, il propose de réfléchir sur des dispositifs expérimentaux qui permettent une relative validation ou invalidation au niveau des résultats produits. Quoique, n'en parlant pas, il n'ignore cependant pas que des analyses comparatives d'applications semblables seraient aussi les bienvenues.

Eléments (provisoires) d'auto-bibliographie

A. Lelu, M. Hallab, F. Papy, S. Bouyahi, H. Rhissassi, N. Bouhaï, H. He, C. Qi, I. Saleh : « Projet NeuroWeb : un moteur de recherche multilingue et cartographique » - Actes de H2PTM'99, coord. J.P. Balpe, S. Natkin, A. Lelu, I. Saleh , Hermès, Paris, 1999.

A. Lelu, S. Aubin « Vers un environnement complet de synthèse statistique de contenus textuels Neuronav version 2 ». Présentation au séminaire ADEST du 13/11/2001 www.upmf-grenoble.fr/adest/seminaires.

A. Lelu, S. Aubin « Vers un environnement complet de synthèse statistique de contenus textuels : Neuronav V3 » – Démonstration industrielle, EGC'2001, coord. IRIN, Nantes, 19/1/2001.

Alain Lelu : « Analyse en composantes locales et graphes de similarité entre textes ». Actes de JADT 2004, G. Purnelle ed., Université catholique de Louvain., 10-12 mars 2004.

B. Hufschmitt : « La philosophie comme multi-, poly-, pan-terminologie, conférence faire au colloque : Termino 2004, Lyon, 23 janvier 2004. http://www.univ-lyon3.fr/partagedessavoirs/termino2004/

B. Hufschmitt : « Terminologie et multi-terminologie dans les écrits argumentatifs en langue naturelle », conférence faire au congrès de l'Isko, Grenoble, 4 juillet 2003. http://isko2003.iut2.upmf-grenoble.fr/

B. Hufschmitt : « Une indexation classique en philosophie rest-elle viable » in Documentation et philosophie II. A propos de l'indexation discursive autour des travaux de Muriel Amar, textes réunis et présentés par Benoit Hufschmitt, Jean-Pierre Cotten et Marie-Madeleine Varet, Presses Universitaires Franc-Comtoises, Besançon, 2003.

B. Hufschmitt : « Thesaurus philosophique, cas de la pensée cartésienne » in Centre de Documentation et Bibliographie philosophiques (Jean-Pierre Cotten dir.): *Outils documentaires pour philosophes, thesaurus, indexation, abstracts*, Annales Littéraires de l'Université de Franche-Comté, diffusion Les Belles Lettres, Paris, 1996.