MASTER EN SCIENCES INFORMATIQUES

Comparaison de logiciels d'analyse de texte

Lattinne, Elisabeth

Award date:
2008

Facultés Universitaires Notre-Dame de la Paix, Namur
Institut d'informatique.
Année académique 2006-2007

Comparaison de logiciels d'analyse de texte

Lattinne Elisabeth

Mémoire présenté en vue de l'obtention du grade de licencié en informatique.

Résumé

De nombreux chercheurs sont confrontés au besoin d'analyser d'immenses corpus de textes. La réalisation de ces analyses se fait la plupart du temps manuellement. Pourtant, des outils existent qui pourraient décharger ces chercheurs de tout ou partie du travail d'analyse. L'objet de ce mémoire est de présenter ces logiciels, les approches conceptuelles qu'ils traduisent ainsi que les attentes des chercheurs à leur égard. Il permettra aux chercheurs de comparer les différents outils d'analyse en regard du corpus à analyser et du type d'analyse souhaité.

Analyse de texte, analyse de contenu, logiciel d'analyse de texte, lexicométrie, analyse des coocurrences, analyse socio-sémantique, analyse cognitivo-discursive, exigences, comparaison, confrontation.

Abstract

Many researchers have to face the challenge of analyzing huge text corpora. This work is most of the time performed manually. Nevertheless, tools are available and might help researchers in the partial or total achievement of such analyses. The aim of the present work is to describe text analysis softwares and the conceptual approaches they feature, as well as the researchers' expectations from such tools. He will allow the researchers to compare the various tools of analysis compared to the corpus to be analyzed and to the wished type of analysis.

Textual analysis, content analysis, text analysis software, lexicometry, co-words analysis, socio-semantic analysis, cognitive-discursive Analysis, requirements, comparison, confrontation.

Avant-propos

Dans le cadre de ce mémoire, je remercie Madame Claire Lobet, mon promoteur, pour son soutien pendant cette année, ses conseils et ses remarques constructives.

Je remercie également Madame Devos, et Messieurs Petit, Heymans et Mayer pour avoir pris le temps de se prêter au jeu des interviews.

Je remercie enfin toutes les personnes qui m'ont soutenue pendant ces trois années et sans lesquelles je ne serai pas arrivée à la réalisation de ce mémoire.

Table des matières

Glossaire

Cluster[1] **:** liste ordonnée de mots apparaissant autour d'un terme de recherche.

Collocation[2] : enclenchement de deux types de lemme ou de mot qui se rencontrent de manière fixe et systématique créant ainsi un concept unitaire et précis.

Concordance[3] : liste des contextes ou ensemble des passages d'un texte où figure un vocable.

Cooccurrence[4] **:** présence simultanée, mais non forcément contiguë, dans un fragment de texte (séquence, phrase, paragraphe, voisinage d'une occurrence, partie du corpus, etc.) des occurrences de deux formes données (ex. : dans la phrase *Le garçon joue*, on dira que *garçon* a pour co-occurrents *le* et *joue.*)

Corpus[5] : ensemble de textes réunis à des fins de comparaison ; servant de base à une étude quantitative ou qualitative.

Co-texte ou contexte[6] : le co-texte d'un mot est l'ensemble des mots qui constituent son entourage qu'ils apparaissent avant ou après dans l'énoncé. Par exemple, dans la phrase *"Une légère pente aboutissait à un fond accidenté.",* le co-texte de *'pente'* est constitué des mots *"une", "légère", "aboutissait", "à", "un", "fond", "accidenté".*

Dictionnaire[7] : Recueil des mots d'une langue ou d'un domaine de l'activité humaine, réunis selon une nomenclature d'importance variable et présentés généralement par ordre alphabétique, fournissant sur chaque mot un certain nombre d'informations relatives à son sens et à son emploi et destiné à un public défini.

Focus group (groupe d'étude)[8] : méthode de consultation des usagers de services destinée à évaluer leurs besoins et leur perception de ces services.

Forme canonique[9] : la forme canonique d'un mot est la forme de ce mot telle qu'on peut la trouver comme entrée d'un dictionnaire par opposition à la forme fléchie.

Forme fléchie[10] : les mots sous forme fléchie comportent un radical et une ou plusieurs désinences. Les désinences sont les morphèmes porteurs des indications de nombre et de genre pour les noms, adjectifs et déterminants, de personnes, de temps et de mode pour les verbes. Ainsi, "lisions" est constitué du radical *lis-* issu de l'item *'lire',* de la désinence temporelle 'i' et de la désinence personnelle *–ons.*

Lemmatisation[11] **:** opération consistant à regrouper les formes occurrentes d'un texte ou d'une liste sous des adresses lexicales. La première étape concerne le regroupement des formes fléchies sous la forme type leur servant d'adresse lexicale ou lemmatisation à proprement parler. La seconde étape consiste en la séparation des formes servant d'adresses lexicales quand elles sont homographes (ex : voile, s.m., et voile, s.f.).

---

[1] http://www.antlab.sci.waseda.ac.jp/
[2] A. M. Scanu, « Hyperbase – Un logiciel pour l'analyse textuelle », www.rilune.org/dese/tesinepdf/**Scanu/Scanu**_Litt%E9ratureetinformatique.pdf
[3] http://fable.ato.uqam.ca/guidexpert-ato///geadoc-vocabu.asp
[4] B. Fracchiolla, A. Kuncova, A. Maisondieu, « Manuel d'utilisation », http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/manuels.htm
[5] B. Fracchiolla, A. Kuncova, A. Maisondieu, « Manuel d'utilisation », http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/manuels.htm
[6] http://www.lirmm.fr/~schwab/pmwiki/pmwiki.php?n=Recherche.Glossaire
[7] http://atilf.atilf.fr/dendien/scripts/tlfiv5/advanced.exe?8;s=4228110375;
[8] http://www.socialeurope.com/mandiv/fr/focus_group.html
[9] http://www.lirmm.fr/~schwab/pmwiki/pmwiki.php?n=Recherche.Glossaire
[10] http://www.lirmm.fr/~schwab/pmwiki/pmwiki.php?n=Recherche.Glossaire
[11] http://mist.univ-paris1.fr/logiciel/def.htm

Lemme[12] : l'unité autonome constituante du lexique d'une langue.

Lexème[13] : morphème lexical d'un lemme, c'est-à-dire une unité de sens et de son qui n'est pas fonctionnelle ou dérivationnelle.

N-gramme[14] : sous-séquence de *n* éléments construite à partir d'une séquence donnée. Les unités concernées sont le plus souvent des caractères ou des mots.

Occurrence[15] : Apparition d'une unité linguistique dans le discours; *p. méton.*, cette même unité, ou nombre de fois où un mot apparaît dans un texte.

Segment répété[16] : suite de deux ou plusieurs mots présents au moins deux fois dans le texte.

Stop-word[17] : mot ignoré lors d'une requête dans les outils de recherche, car son utilisation n'améliore en rien la pertinence des résultats, dans la mesure où il est trop souvent utilisé.

Synapsie : locution dont les éléments sont reliés par des rapports de subordination, comme *planche à repasser, moulin à (moudre le) café*``[18]

TALN[19] **(Traitement Automatique du Langage naturel (ou des langues naturelles)** : domaine d'étude des techniques d'analyse (compréhension) et de génération (production) automatiques d'énoncés oraux ou écrits.

Thésaurus[20] : langage documentaire fondé sur une structuration hiérarchisée d'un ou plusieurs domaines de la connaissance et dans lequel les notions sont représentées par des termes d'une ou plusieurs langues naturelles et les relations entre notions par des signes conventionnels.

Veille[21] : pratique qui englobe les actions de collecte, analyse et diffusion des informations en vue de rendre plus intelligible l'environnement de l'entreprise.

[12] http://fr.wikipedia.org/wiki/Lemme_%28linguistique%29
[13] http://fr.wikipedia.org/wiki/Lex%C3%A8me
[14] http://fr.wikipedia.org/wiki/N-gramme ; http://www.atala.org/AtalaPedie/index.php?title=N-gramme
[15] http://atilf.atilf.fr/dendien/scripts/tlfiv5/visusel.exe?122;s=2790869175;r=4;nat=;sol=1;
http://www.dicodunet.com/definitions/referencement/occurrence.htm
[16] http://www.image.cict.fr/index_alceste.htm
[17] http://www.olf.gouv.qc.ca/ressources/bibliotheque/dictionnaires/Internet/fiches/8361002.html
[18] http://atilf.atilf.fr/dendien/scripts/tlfiv5/advanced.exe?8;s=852305115
[19] http://www.lirmm.fr/~schwab/pmwiki/pmwiki.php?n=Recherche.Glossaire
[20] http://atilf.atilf.fr/dendien/scripts/tlfiv5/advanced.exe?33;s=4228110375;
[21] http://mist.univ-paris1.fr/logiciel/def.htm

Introduction

« Les paroles s'envolent, mais les écrits restent. »

Tel pourrait être le point de départ de ce mémoire.

L'activité de la recherche, aussi diverse soit-elle, est confrontée à un moment ou à un autre à la nécessité de se pencher sur du texte.
Les objectifs poursuivis peuvent être divers : acquérir les concepts théoriques qui sous-tendent l'objet de la recherche, confronter le contenu desdits textes aux acquis théoriques, voire élaborer de nouvelles théories sur base dudit contenu. Les exemples sont légions et nous ne saurions les passer tous ici en revue.

Le sujet de ce mémoire est issu d'une demande de chercheurs pour des outils pouvant les aider à analyser plus efficacement de grands corpus de textes.

La nature et la forme de ces textes peuvent être très variées (interviews, documents officiels, discours politiques, articles scientifiques, réponses à un questionnaire, etc.).

Jusqu'à présent, la plupart des chercheurs étudient ces textes à la main, par ignorance ou par méfiance des outils. Même si certains logiciels existent, ils semblent peu utilisés.

Ces outils permettraient pourtant de les aider et de les décharger de certaines parties de l'analyse.

L'objet de ce mémoire est donc de procéder à une analyse comparative de différents logiciels d'analyse de texte, et ainsi d'offrir au monde de la recherche un aperçu des instruments existants afin de guider leur choix. Ce mémoire permettra donc aux chercheurs de comparer les différents outils d'analyse en regard du corpus à analyser et du type d'analyse souhaité.

Le point de départ de ce mémoire est une liste de logiciels qualifiés de « logiciels d'analyse de texte ». Peu familière de ces outils, un premier travail d'état de l'art a été mené sur ceux-ci. Cet état de l'art a permis de comprendre que sous le vocable « analyse de texte » se cachaient en fait plusieurs approches conceptuelles ou écoles de pensée. Nous avons donc investigué ces différentes approches. En parallèle, la liste des logiciels a été complétée et a permis d'affiner les concepts.

Ces allers-retours entre théorie et logiciels ont posé les limites de ce mémoire. Nous nous sommes en effet intéressés à une certaine catégorie de logiciels, alors que d'autres ont été volontairement exclus car ne répondant pas aux approches conceptuelles identifiées.

Il est évidemment en soi très intéressant d'identifier des logiciels d'analyse de texte ainsi que les outils qu'ils offrent. Encore faut-il que ces outils présentent une utilité aux personnes du terrain. Il apparaît donc essentiel de cerner les attentes à ce sujet. Pour ce faire, nous avons procédé à des interviews de chercheurs confrontés régulièrement à des matériaux textuels dans leurs activités de recherche. L'analyse de ces interviews a permis de dégager leurs attentes et de confronter celles-ci aux outils existants.

Les points suivants constitueront donc les chapitres de ce mémoire. Ces chapitres ne correspondent pas exactement à la chronologie du travail effectué. En effet, la réalisation des interviews est survenue assez rapidement afin d'aider à délimiter les approches à retenir dans ce mémoire parmi les très nombreuses approches existantes en analyse de texte.

Dans un premier temps (Partie 1), il s'agira de se familiariser avec la notion d'analyse de texte, ainsi qu'avec les méthodes d'analyse qui ont fait l'objet d'une traduction informatique.

En ce qui concerne la notion d'analyse de texte (chapitre 1), il s'agira de comprendre exactement ce que recouvre ce concept, le(s) but(s) qu'il poursuit et les méthodes permettant d'atteindre ces buts.

Certaines de ces méthodes ont fait l'objet d'une traduction informatique (Chapitre 2). Nous nous pencherons sur celles-ci et sur leurs caractéristiques.

Dans un deuxième temps (Partie 2), nous procéderons à la présentation des logiciels identifiés, à la recherche des exigences des chercheurs en la matière et à la confrontation de celles-ci aux outils analysés.

En ce qui concerne les logiciels (Chapitre 3), il s'agira de présenter leurs caractéristiques et d'identifier la (les) méthode(s) supportée(s), leurs domaines d'application ainsi que les fonctionnalités qu'ils offrent en matière d'analyse de texte.
De plus, certains outils seront sélectionnés et analysés plus en profondeur sur base de critères déterminés préalablement.
Enfin, ils seront également comparés entre eux lorsque cela est possible et évalués quantitativement.

L'analyse des interviews (Chapitre 4) permettra de découvrir les objectifs poursuivis en matière d'analyse de texte ainsi que les méthodes utilisées pour les atteindre. Elle aura également pour objectif de mettre en lumière des critères d'évaluation des logiciels retenus.

Enfin, le dernier chapitre (Chapitre 5) aura pour objectif de déterminer si les logiciels identifiés peuvent répondre aux attentes des chercheurs. Sur base de critères et de caractéristiques mis à jour tout au long du mémoire, il s'agira de déterminer si et dans quelle mesure ces logiciels peuvent suppléer le chercheur dans certaines des étapes de l'analyse.

Nous terminerons par quelques éléments de conclusions.

Partie 1 : L'analyse de texte et ses méthodes

La première partie de ce mémoire va être consacrée à l'analyse de texte et aux méthodes sur lesquelles elle se base, méthodes qui ont parfois fait l'objet d'une traduction informatique.

Le premier chapitre va être consacré à la notion d'analyse de texte et aux différents aspects qu'elle peut recouvrir. En effet, au cours des recherches effectuées dans le cadre de ce mémoire, il est rapidement apparu que l'activité d'analyse de texte est diversifiée et influencée par ses utilisateurs.

Tel que cela sera expliqué dans le chapitre, nous nous sommes centrés sur une forme d'analyse de texte particulière : l'analyse de contenu. Celle-ci peut elle-même être divisée en différents types d'analyse qui seront présentés

L'analyse de texte ne suppose pas par elle-même l'utilisation de l'informatique. Elle est, la plupart du temps, effectuée manuellement suivant une méthode générale présentée en annexe (Annexe 1), par manque d'outils par le passé, mais également par manque de connaissance de ceux-ci, voire par méfiance actuellement.

Cette manière de faire a toutefois montré certaines limites (masse des données, temps, complétude, objectivité, etc.). Il est dès lors paru intéressant de créer des outils permettant de contrer certains problèmes. Des logiciels ont été mis au point afin d'aider le chercheur dans sa démarche. Ceux-ci ont essayé de formaliser différentes méthodes mises au point pour l'analyse de texte.

Des outils ont donc été développés afin de se substituer à l'analyste dans certaines étapes de sa recherche.

Le deuxième chapitre sera donc consacré aux méthodes d'analyse de texte qui ont été informatisées.

Une remarque préalable s'impose avant d'aborder ces notions. Dans la démarche d'analyse de texte, il convient de bien distinguer le but poursuivi et la méthode utilisée. Lorsqu'un chercheur analyse un texte, il poursuit un but, un objectif, souvent en rapport avec l'objet de sa recherche. Il lui appartient donc de le définir précisément.
Par ailleurs, ce même chercheur se doit de déterminer la méthode qu'il suivra pour atteindre ce but. Il arrive en effet fréquemment qu'un même objectif puisse être atteint par différentes méthodes et il appartient au chercheur de faire un choix.

Cette remarque trouve à s'appliquer dans le cadre de cette première partie. Les types d'analyse de texte identifiés dans le premier chapitre sont plutôt vus comme des buts à atteindre, tandis que les méthodes informatisées seront plutôt considérées, comme leur nom l'indique, comme les méthodes permettant d'atteindre les objectifs du premier chapitre.

Bien évidemment, rien n'étant jamais totalement noir ou blanc, certaines nuances seront parfois apportées.

Chapitre 1 : L'analyse de texte

Le premier chapitre de cette partie va être consacré à l'analyse de texte et à ses différentes variantes

Des sources consultées dans le cadre de ce mémoire, il ressort que l'analyse de texte peut intervenir dans de nombreuses disciplines : sociologie, psychologie, journalisme, linguistique, littérature, sciences politiques, sciences documentaires, ethnologie, histoire, psychiatrie, anthropologie, marketing, etc.

Si l'on examine les domaines cités, on peut toutefois remarquer qu'ils tournent principalement autour des sciences humaines entendues au sens large, économiques et littéraires.
D'autres domaines de type technique ou « sciences pures », tels l'ingénierie, la biologie, la chimie, l'informatique, etc., semblent ignorer cette démarche.

Il sera donc intéressant dans le cadre de ce mémoire de déterminer si le domaine de l'informatique est ou pourrait être concerné par l'analyse de texte.

Le terme « texte » est à prendre au sens large. Il peut s'agir d'un document écrit produit directement par son auteur, manuellement ou à l'aide de l'informatique, mais également de productions orales, interviews, discours, retranscrites par la suite ; le point commun de ces textes étant de pouvoir être qualifiés de « communication », c'est-à-dire de « transport de significations d'un émetteur à un récepteur, contrôlé ou non par celui-là ».[22]

Ces communications peuvent être plus ou moins structurées. Cela peut aller d'une simple liste de mots fournie par des interviewés lors d'une recherche portant sur des stéréotypes, par exemple, à un texte élaboré avec minutie dans le cadre d'un article pour une revue scientifique, en passant par des textes « semi-structurés » dans le cadre d'entretiens sur base de questions ouvertes.

Deux grands courants théoriques ont été dégagés en matière d'analyse de texte : l'analyse de contenu et l'analyse de discours.

L'analyse de discours tient une place particulière dans l'analyse de texte. [23]

Il s'agit d'une approche « *multidisciplinaire qualitative et quantitative qui étudie le contexte et le contenu du discours oral ou écrit* ». [24]

Cette discipline s'est développée dans les années soixante en France, en Grande-Bretagne et aux Etats-Unis.

Elle a emprunté des concepts à de nombreuses autres disciplines : sociologie, philosophie, psychologie, informatique, communication, linguistique et histoire.

Elle s'intéresse « *aux concepts, à la linguistique et à l'organisation narrative des discours oraux et écrits qu'elle étudie* ». [25]

[22] L. Bardin, L'analyse de contenu, 11° édition, Paris, 2003, P.U.F., p. 36.
[23] http://fr.wikipedia.org/wiki/Analyse_du_discours
[24] http://fr.wikipedia.org/wiki/Analyse_du_discours

L'analyse du discours a pour unité d'analyse « *la proposition ou l'énoncé, permettant ainsi de développer une étude des composants de phrase d'une part, et de réintroduire l'énonciateur et la situation dans l'analyse d'autre part* ».[26]

Certains[27] placent l'analyse du discours parmi les méthodes d'analyse de contenu, d'autres[28] l'en distinguent.

« *…l'analyse de discours présume l'existence d'une réalité, existant dans l'énoncé, formée à travers l'argumentation, la stylistique, la forme et les enchaînements du discours oral ou écrit* ». [29] C'est en cela qu'elle se différencierait de l'analyse de contenu.

« *Pour l'analyse de contenu, le discours est un reflet de la réalité, pour l'analyse de discours, celui-ci constitue la réalité en soi.* »[30]

L'analyse de discours étant un courrant particulier de l'analyse de texte, ce mémoire ne traitera que du courant de l'analyse de contenu, plus large.

Le premier point de ce chapitre aura donc pour objectif d'identifier les différents types d'analyse de contenu qui existent en sciences humaines, berceau de l'analyse de texte.

Un deuxième point présentera une synthèse de la matière.

1. L'analyse de contenu[31]

L'analyse de contenu est définie traditionnellement comme « *une technique de recherche pour la description objective, systématique et quantitative du contenu manifeste des communications, ayant pour but de les interpréter* ».[32]

Cette définition étant restrictive, Bardin en propose une plus large. L'analyse de contenu vise « *un ensemble de techniques d'analyse des communications visant, par des procédures systématiques et objectives de description du contenu des messages, à obtenir des indicateurs (quantitatifs ou non) permettant l'inférence de connaissances relatives aux conditions de production/réception (variables inférées) de ces messages* ».[33]

Elle fait partie des méthodes d'analyse qualitative[34] des sciences humaines et sociales.[35]

Elle a pour objectif de « *comprendre les activités cognitives du locuteur. (Son idéologie, ses attitudes…)* ». [36]

---

[25] http://fr.wikipedia.org/wiki/Analyse_du_discours
[26] R. Ghiglione et al., L'analyse automatique des contenus, Paris, Dunod, 1998, p. 1.
[27] http://phnk.com/files/m2-teq-bardin-livret.pdf
[28] http://fr.wikipedia.org/wiki/Analyse_du_discours
[29] http://fr.wikipedia.org/wiki/Analyse_du_discours
[30] http://fr.wikipedia.org/wiki/Analyse_du_discours
[31] M.-N., Schurmans, « Introduction aux démarches compréhensives », 12 mai 2004, http://www.unige.ch/fapse/SSE/teachers/schurmans/notesCours12%20mai.rtf ; N. Patanella, « Note sur la méthodologie de la science politique - A *destination des étudiants du dea en relations internationales et intégration européenne* », http://www.ulg.ac.be/polgereg/Publications/Methodo.pdf ; C. Suter, « L'analyse de contenu », Sociologie générale I : Introduction à la sociologie, http://www.unine.ch/socio/enseignement/socio1_2004/soc1_8_6.doc
[32] Berelson (1948) in L. Renaud, « Méthodes de recherche en communication », mars 2003, http://www.er.uqam.ca/nobel/r13761/medias/cours10.pdf et in L. Bardin, o.c., p. 21.
[33] L. Bardin, o.c., p. 47.
[34] Autres méthodes : les « focus groups », l'observation et la « participation » ; http://fr.wikipedia.org/wiki/Analyse_de_contenu
[35] http://fr.wikipedia.org/wiki/Analyse_de_contenu

L'analyse de contenu s'intéresse à ce qui est caché dans le texte.[37] Il s'agit de l'interpréter.[38] L'idée est de dire non à l'« illusion de la transparence » des faits sociaux. Il n'y a pas de compréhension spontanée. Il est donc nécessaire d'avoir des outils appropriés afin de découvrir ce qui se cache derrière le texte.[39]

La visée de l'analyse de contenu n'est pas purement descriptive. Elle a aussi pour fonction « *l'inférence*[40] de connaissances relatives aux conditions de production (ou éventuellement de réception), à l'aide d'indicateurs (quantitatifs ou non) ».[41]

La démarche de l'analyse de contenu suit donc le trajet suivant :

Description → inférence → interprétation

La question principale de l'analyse de contenu est selon Lasswell "*Qui dit quoi à qui et avec quel effet ?*".[42]

Elle peut s'appliquer à toute communication langagière (articles, déclarations politiques, conversations, etc.), mais aussi à des supports visuels (photographies, films, etc.).

Elle doit être distinguée de l'analyse linguistique qui a pour « objectif de comprendre le fonctionnement du langage en tant que tel ».[43]

Il existe différents types d'analyse de contenu, regroupés de manières parfois différentes selon les auteurs. Le présent regroupement se base sur ceux de Laurence Bardin,[44] Paul Revollon et Claude Larrecq.[45]

Ils identifient deux grandes familles d'analyses de contenu : les analyses thématiques et les analyses formelles.

1.1. Les analyses thématiques

Les analyses thématiques « *tentent de mettre en évidence les représentations sociales ou les jugements des locuteurs à partir d'un examen de certains éléments constitutifs du discours* ».[46]

Les analyses thématiques s'intéressent donc au fond du texte. Elles visent à identifier les thèmes abordés et la manière dont l'auteur en parle.

En font partie l'analyse catégorielle et l'analyse de l'évaluation.

---

[36] P. Revollon et C. Larrecq, « L'analyse de contenu », http://www.inh.fr/enseignements/idp/idp2005/outils/etude_marche/contenu_psycho_socio.pdf
[37] L. Bardin, o.c., p. 13.
[38] L. Bardin, o.c., p. 16.
[39] L. Bardin, o.c., p. 31.
[40] « opération logique, par laquelle on admet une proposition en vertu de sa liaison avec d'autres propositions déjà tenues pour vraies » ; L. Bardin, o.c., p. 43, note 2.
[41] L. Bardin, o.c., p. 43.
[42] C. Suter, o.c.
[43] P. Revollon et C. Larrecq, o.c.
[44] L. Bardin, o.c.
[45] P. Revollon et C. Larrecq, o.c. ; http://phnk.com/files/m2-teq-bardin-livret.pdf
[46] P. Revollon et C. Larrecq, o.c.

a) L'analyse catégorielle[47]

La première forme d'analyse thématique est l'analyse catégorielle qui recouvre différents sens selon la discipline dans laquelle elle est employée.

Elle est en premier lieu sollicitée dans le cadre de l'analyse grammaticale.

En ce qui concerne les mots, « l'analyse catégorielle consiste à spécifier la *nature* et *l'espèce* des mots et des propositions subordonnées. »[48] Pour les noms variables (noms, pronoms, verbes, articles, adjectifs), on spécifie également leur *forme*.

« Spécifier la nature d'un mot, c'est l'attribuer à « sa » catégorie lexicale (ou classe de mots ou, en termes traditionnels, « partie du discours »). »[49]

Les catégories lexicales sont : nom, verbe, adjectif, etc.

« Spécifier l'espèce d'un mot, c'est l'attribuer à une sous-catégorie (ou sous-classe). »[50]

Par exemple, pour les noms, on distingue les sous catégories *commun* et *propre*, pour les articles *défini*, *indéfini* et *partitif*.

La spécification de la forme comprend le nombre, le genre, la personne et le mode.

« L'analyse catégorielle des propositions subordonnées consiste à en spécifier l'espèce.
D'après la nature du mot qui introduit la proposition subordonnée, on distingue :
- proposition ***relative***,
- proposition ***conjonctive*** et
- proposition ***interrogative*** (ou *exclamative*) ***indirecte***.
S'y ajoutent les propositions ***infinitives*** et ***participes***, qui ne sont pas introduites et qui tirent leur nom de la forme du verbe prédicatif. »[51]

Comme nous le verrons par la suite, cette forme d'analyse catégorielle est intéressante dans le cadre de la construction de dictionnaires ou thésaurus équipant certains logiciels d'analyse de texte. Ces dictionnaires et thésaurus sont utilisés dans le cadre de l'indexation du texte. Nous reviendrons sur ces notions par après.

L'analyse catégorielle est également utilisée en sciences humaines comme méthode d'analyse de contenu.

Elle peut alors être de deux types :

- Lexicale : « Analyse du nombre, fréquence, récurrences, proximité de mots ou d'expression »[52] ;
- Thématique : « Analyse des occurrences, cooccurrences ou l'ordre d'apparition des unités de signification (thème) » [53] ; elle « consiste à calculer et à comparer les

[47] B. Schwischay, "Mémento d'analyse grammaticale", 3 mars 2002, http://www.home.uni-osnabrueck.de/bschwisc/archives/analyse.pdf ; L. Bardin, o.c., p. 40.
[48] B. Schwischay, o.c., p. 4.
[49] B. Schwischay, o.c., p. 4.
[50] B. Schwischay, o.c., p. 4.
[51] B. Schwischay, o.c., p. 10.
[52] http://www.er.uqam.ca/nobel/r32700/Cours%20site/SITE%20%20COM7103/cueilletteetanalyse.htm

> fréquences de certains éléments (le plus souvent les thèmes évoqués) et à les regrouper en catégories significatives. L'hypothèse est qu'une fréquence élevée d'une idée = cette idée est importante pour le locuteur. La démarche est essentiellement quantitative. »[54]

En sciences humaines, l'analyse catégorielle a donc pour objectif d'identifier l'objet du texte. De quoi parle-t-il ? Quels sont les thèmes abordés ?

Dans la définition reprise, la méthode utilisée est quantitative. Chaque unité lexicale fait l'objet d'un décompte. L'identification d'un thème correspondra à une fréquence importante.

La différence présentée entre « analyse lexicale » et « analyse thématique » reflète certaines questions qui se posent quant au décompte des unités lexicales. Que doit-on compter ? Et comment ?

L'analyse lexicale établit un décompte au niveau le plus simple, les unités lexicales, les mots. En soi, ce terme pose déjà question. Que recouvre-t-il ? Les mots simples, les mots composés, les expressions ? Faut-il appliquer une lemmatisation ? Comment identifier les termes ambigus ?

Un exemple : « Je livre un livre. »
Cet exemple nous montre un problème d'ambiguïté. Il est important de ne pas compter ensemble les deux « livre », l'un étant un verbe et l'autre un substantif.

L'analyse thématique suppose un regroupement de certaines unités sémantiquement proches formant un thème. Ce regroupement est en partie subjectif, propre à chaque personne.

Un exemple : « Le chat mange la souris. Le poisson tourne dans son bocal. »
Devons-nous opérer un regroupement entre certains termes ? Devons-nous considérer que « chat », « souris » et « poisson » font partie du thème « animal » et que le thème de ce (mini) texte est « animal » ? Ou devons-nous séparer les thèmes « chat » et « poisson » (voire « souris ») ?
Le texte est évidemment fort court pour trancher, mais il n'a qu'une visée explicative.

Nous verrons par la suite que la méthode quantitative utilisée pour identifier des thèmes n'est pas soutenue unanimement et que d'autres méthodes ont été développées de type qualitatif.

b) L'analyse de l'évaluation

La deuxième forme d'analyse thématique est l'analyse de l'évaluation. Celle-ci « *porte sur les jugements formulés par le locuteur. La fréquence des différents jugements est calculée mais aussi leur direction (jugement positif ou négatif) et leur intensité (ex. analyse d'un jugement à la cour)* ».[55]

L'objectif est donc de repérer les jugements exprimés par l'auteur. Il ne s'agit plus d'identifier un « quoi » mais un « comment ».

---

[53] http://www.er.uqam.ca/nobel/r32700/Cours%20site/SITE%20%20COM7103/cueilletteetanalyse.htm
[54] P. Revollon et C. Larrecq, o.c.
[55] L. Renaud, o.c.

La méthode utilisée est le découpage du texte en unités de signification et l'attribution d'une charge évaluative à celles-ci. Tout le texte n'est toutefois pas pris en considération. Seules les propositions exprimant une évaluation sont gardées.[56]

Ce type d'analyse peut être fait en complément de l'analyse catégorielle. Après avoir identifié les thèmes d'un texte, il peut être intéressant de voir la manière dont l'auteur parle de ces thèmes. Dans ce cadre, seules les propositions exprimant un jugement seront sélectionnées.

Exemple : « La politique suivie par ce Ministre a eu des conséquences désastreuses en matière économique »
Cette phrase montre un exemple de jugement et peut donc être retenue pour une analyse de l'évaluation.

1.2. Les analyses formelles

Les analyses formelles « *portent sur les formes de l'enchaînement du discours et mettent l'accent sur la manière dont les éléments du message sont agencés* ».[57]

Contrairement à l'analyse thématique qui s'intéressait au fond d'un texte, les analyses formelles se penchent plutôt sur la forme.

Elles se subdivisent en différents types : l'analyse de l'expression, l'analyse des relations et l'analyse de l'énonciation.[58]

a) L'analyse de l'expression

Le terme « expression » peut prendre différents sens :
- Action de rendre manifeste par toutes les possibilités du langage, plus particulièrement par celles du langage parlé et écrit, ce que l'on est, pense ou ressent[59]
- Ensemble des signifiants[60] (quelle qu'en soit la substance phonique ou graphique) par opposition à contenu ou ensemble des signifiés[61]

Ces deux notions permettent d'éclairer la définition de l'analyse de l'expression.

Celle-ci « *porte sur la forme de la communication, qui donne des informations sur l'état d'esprit du locuteur et ses dispositions idéologiques (vocabulaire, longueur des phrases, ordre des mots, hésitations…)* ».[62]

L'objectif est d'en apprendre plus sur l'auteur du texte aux travers des aspects formels de celui-ci.

Les indicateurs retenus seront entre autres : des indicateurs lexicaux et stylistiques, les enchaînements logiques, les agencements séquentiels, la structure narrative, etc.[63]

---

[56] L. Bardin, o.c., p. 209.
[57] P. Revollon et C. Larrecq, o.c.
[58] P. Revollon et C. Larrecq, o.c.
[59] http://atilf.atilf.fr/dendien/scripts/tlfiv5/saveregass.exe?17;s=802215720;r=1;;
[60] « Partie formelle, matérielle et sensible du signe », http://atilf.atilf.fr/dendien/scripts/tlfiv5/advanced.exe?48;s=802215720;
[61] http://atilf.atilf.fr/dendien/scripts/tlfiv5/saveregass.exe?17;s=802215720;r=1;;
[62] P. Revollon et C. Larrecq, o.c.
[63] L. Bardin, o.c., p. 255.

Un domaine particulièrement concerné par l'analyse de l'expression est la vérification de l'authenticité de documents. En effet, chaque personne a son propre style, utilise les mêmes tournures de phrases, voire les mêmes expressions, et l'analyse de ceux-ci doit permettre de vérifier l'origine du texte.

b) L'analyse des relations

L'analyse des relations porte, comme son nom l'indique, sur les relations existant entre certains éléments du texte.

Elle peut prendre différentes formes : analyse des cooccurrences, analyse structurale et analyse du récit.

b.1) Analyse des cooccurrences

L'analyse des cooccurrences « *examine les associations de thèmes dans les séquences de la communication (on observe des présences simultanées de deux ou plusieurs éléments dans une même unité de contexte). Cela est censé informer le chercheur sur des structures mentales et idéologiques ou sur des préoccupations latentes* ».[64]

L'objectif de ce type d'analyse est d'identifier les thèmes fortement associés dans l'esprit de l'auteur. Cette analyse présente des liens avec l'analyse thématique abordée auparavant puisqu'il s'agit aussi d'identifier des thèmes. Elle s'en distingue toutefois dans la mesure où il s'agit de déterminer si des thèmes sont liés.

La méthode utilisée est de voir si des thèmes sont simultanément présents dans une même unité de contexte, la question étant de définir correctement cette dernière. Cela suppose un découpage du texte, susceptible d'influencer les résultats.

Si nous reprenons un exemple déjà cité : « Le chat mange la souris. Le poisson tourne dans son bocal. ». On peut décider de prendre comme unité de découpage la phrase. On va donc regarder quels termes sont associés. Ici, « chat » et « souris », « poisson » et « bocal ». Pour que les résultats soient significatifs, il conviendrait que le texte se poursuive et d'examiner si d'autres phrases présentent les mêmes associations, et de compter celles-ci.

b.2) Analyse structurale[65]

Le terme « structural » peut être défini comme « *qui étudie ce qui constitue la/les structure(s) ; qui isole un ensemble d'éléments et de relations formelles pour étudier un phénomène sans faire appel à la signification* ».[66]

L'analyse structurale met « *l'accent sur la manière dont les éléments du message sont agencés pour transmettre l'information* ».[67] Elle vise à « *mettre en évidence les principes qui organisent les éléments du discours, de manière indépendante du contenu même de ces éléments* ».[68]

---

[64] P. Revollon et C. Larrecq, o.c.,
[65] C.-A. Audet, « Méthode d'analyse structurale de la phrase », http://www.aide-doc.qc.ca/le.grammairien/ftp/analstruct.pdf
[66] http://atilf.atilf.fr/dendien/scripts/tlfiv5/advanced.exe?73;s=802215720;
[67] L. Renaud, o.c.
[68] P. Revollon et C. Larrecq, o.c.

« … *on se penche sur l'agencement des différents items, en essayant de découvrir des constantes significatives dans les relations qui organisent ces items entre eux.* »[69]

Comme cela ressort des définitions, on ne s'intéresse pas au fond du texte, mais à sa forme. L'objectif est de trouver la manière dont les différents éléments formels d'un texte se combinent entre eux pour exprimer ce que l'auteur veut communiquer.

Pour ce faire, ce type d'analyse va emprunter des éléments de linguistique et de littérature : les principes d'organisation sous-jacents du texte, les systèmes de relation, les schèmes directeurs, les règles d'enchaînement, d'association, d'exclusion, d'équivalence, les agrégats organisés de mots ou d'éléments de signification, les figures rhétoriques, etc.[70]

Par exemple, on peut s'intéresser aux types de phrase employés : simples, composées, complexes.

Exemples : « Je mange une pomme » : phrase simple
« Je lis un livre et je regarde la télévision » : phrase composée

b.3) Analyse du récit

Le récit est un texte narratif, « c'est-à-dire racontant un événement ou une histoire composée d'une série d'événements ».[71]

Le récit peut être analysé à l'aide de notions comme le schéma narratif, le schéma actantiel, le point de vue, le style direct, le style indirect, le style indirect libre, la focalisation interne, la focalisation externe, etc. [72]

Le récit peut être étudié dans une optique littéraire-linguistique, mais également dans une perspective psychologique ou sociologique, voire dans un but historique.[73]

Il est également possible d'étudier ses conditions de production ou de réception.[74]

c) L'analyse de l'énonciation

L'énonciation est « *à l'intérieur du texte un ensemble de traces qui manifestent l'acte par lequel un auteur a produit ce texte* »[75] ou un « *acte de production linguistique par opposition à énoncé* ».[76]

L'analyse de l'énonciation « *porte sur le discours conçu comme un processus dont la dynamique propre est en elle-même révélatrice. Le chercheur est alors attentif à des données telles que le développement général du discours, l'ordre de ses séquences, répétitions, ruptures du rythme (ex. analyse d'un film)* ».[77]

L'objectif est donc d'identifier la manière dont le texte est produit, le processus qui l'a créé.

---

[69] P. Revollon et C. Larrecq, o.c.
[70] L. Bardin, o.c., p. 278.
[71] http://www.lettres.net/files/recit.html
[72] http://www.lettres.net/files/recit.html
[73] M. Fayol, Le récit et sa construction, Paris, Delachaux et Niestlé, 1985, p. 7.
[74] Y. Reuter, L'analyse du récit, Paris, Dunod, 1997, p. 7.
[75] http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW/manuelsL3/manuel-3.41.pdf
[76] http://atilf.atilf.fr/dendien/scripts/tlfiv5/advanced.exe?33;s=869060655;
[77] L. Renaud, o.c.

Cet aspect « processus » est essentiel. C'est la parole en acte qui est étudiée. « *L'analyse de l'énonciation considère qu'un travail se fait lors de la production de parole, qu'un sens s'élabore, que des transformations s'opèrent* ».[78]

Ce type d'analyse se base sur les structures et les éléments formels du texte.[79]

Au niveau méthodologique, différentes analyses vont se combiner : analyse syntaxique et paralinguistique, analyse logique (agencement du discours), analyse des éléments formels atypiques (omissions, failles logiques, silences, etc.) et des figures de rhétorique, analyse séquentielle, analyse du style.[80]

2. Synthèse

L'objectif de ce point est de donner un aperçu synthétique des différents types d'analyse de contenu présentés ci-dessus.

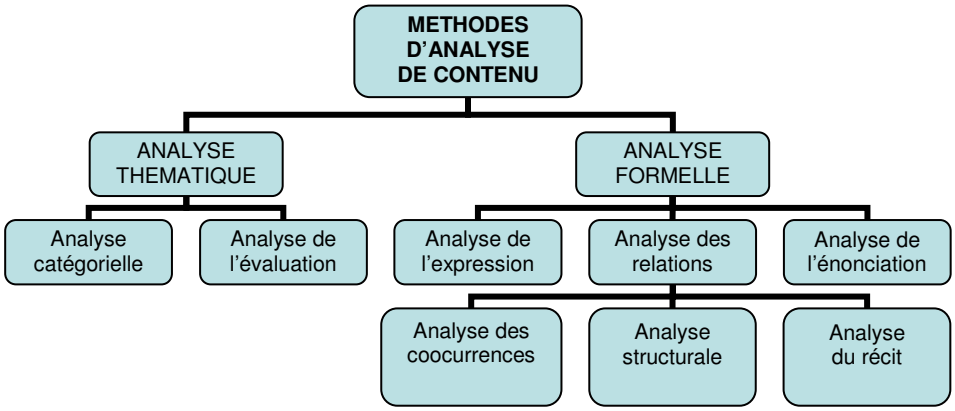
Nous allons proposer un tableau de synthèse en se focalisant sur deux points : le quoi et le comment, l'objectif et la méthodologie. Il convient toutefois de noter que les méthodologies citées ne sont pas exhaustives. En effet, tout chercheur est susceptible de créer une nouvelle méthode afin d'atteindre un des buts cités.

	Quoi ?	**Comment ?**
Analyses thématiques	représentations sociales ou jugements du locuteur	examen de certains éléments constitutifs du discours
▪ **Analyse catégorielle**	thèmes – objets du texte	comptage – quantitatif, une présence élevée est synonyme d'importance
▪ **Analyse de l'évaluation**	jugements	recherche des propositions évaluatives et attribution d'une charge évaluative : fréquence, direction, intensité
Analyses formelles	éléments de forme du texte	formes de l'enchaînement du discours
▪ **Analyse de l'expression**	état d'esprit, dispositions idéologiques de l'auteur	formes de la communication (vocabulaire, longueurs des phrases, ordre des mots, hésitations, etc.) – indicateurs formels (indicateurs lexicaux et stylistiques, enchaînement logique, agencement séquentiel, structure narrative, etc.)
▪ **Analyse des relations**		
- **Analyse des cooccurrences**	thèmes fortement associés	simultanéité de thèmes dans une même unité de contexte

[78] L. Bardin, o.c., p. 224.
[79] L. Bardin, o.c., p. 223.
[80] L. Bardin, o.c., p. 225 et 230.

- **Analyse structurale**	structure du discours, de manière indépendante du contenu	les schèmes directeurs, les règles d'enchaînement, d'association, d'exclusion, d'équivalence, les agrégats organisés de mots ou d'éléments de signification, les figures rhétoriques, etc.
- **Analyse du récit**	différentes perspectives : littéraire-artistique ; psychologique ou sociologique ; historique	le schéma narratif, le schéma actantiel, le point de vue, le style direct, le style indirect, le style indirect libre, la focalisation interne, la focalisation externe, etc.
▪ **Analyse de l'énonciation**	discours comme processus – recherche de la dynamique	discours = processus – développement général, ordre des séquences, répétitions, ruptures du rythme, etc.

Le schéma ci-dessous reprend l'organisation hiérarchique de ces différents types d'analyse.



Comme cela a été dit lors de l'introduction, l'étude des types d'analyse de contenu a mis en évidence la différence qu'il existe entre l'objectif poursuivi et les méthodes à suivre pour atteindre ce but. Seules certaines de ces méthodes ont été présentées, étant donné leur diversité.

Le chapitre suivant va se pencher sur certaines de ces méthodes qui ont été informatisées, mais aussi sur d'autres, informatisées, mais non encore citées.

Chapitre 2. Les méthodes d'analyse de texte informatisées

Après la présentation des différents types d'analyse de contenu, ce deuxième chapitre va se pencher sur certaines méthodes qui ont été transposées dans des logiciels informatiques.

Ces méthodes peuvent évidemment être appliquées manuellement.

Celles-ci sont nombreuses et chaque nouveau logiciel est susceptible d'en créer une nouvelle. Il est cependant possible d'identifier des méthodes récurrentes. Nous allons présenter les principales.

Un premier constat général peut toutefois s'imposer. La majorité, voire toutes les méthodes présentées, ressortent d'une manière ou d'une autre, de l'analyse thématique identifiée dans le premier chapitre, et plus particulièrement de l'analyse catégorielle utilisée dans les sciences humaines. Un des objectifs principaux des méthodes abordées est de permettre au chercheur d'identifier les thèmes du corpus qu'il analyse.

1. Les méthodes quantitatives : de la lexicométrie aux analyses statistiques de données textuelles

Cette première catégorie de méthodes d'analyse ayant fait l'objet d'une informatisation est de type quantitatif.

Il s'agit de se baser sur des chiffres, des décomptes, des calculs statistiques, l'idée sous-jacente étant que cela serait plus « objectif ».

La méthode la plus simple et préalable à d'autres calculs savants consiste à compter les différentes formes présentes dans le corpus. Il s'agit de l'analyse lexicométrique au sens strict.
Par la suite, il est possible d'appliquer d'autres formules statistiques aux premiers résultats obtenus afin de confirmer ceux-ci ou d'en produire de nouveaux.

1.1. L'analyse lexicométrique[81]

Dans son acception la plus simple, l'analyse lexicométrique, ou lexicométrie, « *compare les décomptes réalisés à partir du repérage des occurrences d'unités lexicales (formes, segments, types généralisés, etc.) dans les différentes parties d'un corpus de textes* ».[82]

La lexicométrie a donc pour objectif la mesure du lexique (vocabulaire) d'un corpus.
Cette mesure permet d'identifier des thèmes, l'idée étant que plus la fréquence est grande, plus le thème est important pour le locuteur.

Elle permet d'obtenir la liste des unités lexicales classées par ordre alphabétique ou par ordre de fréquence croissante ou décroissante.
Il est également possible de partitionner le texte afin de procéder à des comparaisons de vocabulaires entre les différentes parties.[83]

[81] http://www.reunion.iufm.fr/Dep/Lettres/_L'approche_lexicométrique
[82] B. Fracchiolla, A. Kuncova, A. Maisondieu, « Manuel d'utilisation », http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/manuels.htm
[83] http://www.weblettres.net/sommaire.php?entree=20&rubrique=75

La lexicométrie suit les étapes suivantes.

Elle dresse d'abord « l'inventaire de toutes les "formes graphiques brutes" (ou "lexèmes", équivalents des "mots") du corpus à analyser, dans un double classement : par ordre alphabétique et par ordre de fréquence d'occurrence ». [84]

Ensuite, elle construit le « Tableau Lexical Entier » (ou T.L.E.) du corpus. Celui-ci est « composé d'autant de lignes (ou de colonnes) qu'il y a de "mots", classés en rang de fréquence décroissante (éventuellement, au-delà d'un certain seuil), et d'autant de colonnes (ou de lignes) qu'on aura préalablement partitionné ce corpus en parties distinctes[85] ». [86]

Différentes variantes du TLE sont possibles :

- « selon la manière de partitionner le corpus pour faire apparaître différentes sortes de variations distributionnelles pertinentes ;
- selon la manière de "traiter" les formes graphiques brutes du corpus, avec effet d'en réduire plus ou moins le nombre - depuis le respect absolu des occurrences telles quelles (= "chaînes de caractères séparés par des délimiteurs", avec toutes les variétés de typographie parasite qu'on rencontre dans la pratique) jusqu'aux multiples options possibles de réduction des variétés non pertinentes (simples corrections orthographiques, conventions ad hoc d'homogénéisation pour les élisions, les locutions, les caractères diacritiques en minuscule/Majuscule, décisions de lemmatisation des formes fléchies, de traitement différé des "mots-outils", etc.). »[87]

Enfin, « les calculs ultérieurs consistent à comparer les « profils lexicaux » (exprimés par les fréquences différentielles des mots, dans les colonnes - ou lignes - du Tableau) des différentes parties du corpus. » [88]

Exemple : « Le référencement sur le moteur de recherche Google des sites Internet des éditeurs de presse quotidienne belge francophone et germanophone, membres de Copiepresse, a repris jeudi, ont annoncé les deux parties dans un communiqué commun. "Cette décision a été prise de commun accord par Google Inc et Copiepresse dans le cadre de la reprise d'un dialogue constructif qui est intervenue entre les deux sociétés", précisent les entreprises dans le communiqué. »[89]

Tableau lexical entier :

forme	**Fréquence**	**Forme**	**Fréquence**
de	6	été	1
le	4	francophone	1
dans	3	germanophone	1
a	2	Inc	1
commun	2	Internet	1

[84] Jacques Jenny, « Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine. Etat des lieux et essai de classification. », Article publié dans le *Bulletin de Méthodologie Sociologique (B.M.S.),* n° 54, mars 1997, p.64-112, http://pageperso.aol.fr/jacquesjenny/ATBMS.htm

[85] par exemple, locuteurs différents (par ex. enquêtes par questions ouvertes), et/ou textes successifs dans le temps (par ex. série chronologique des discours d'une même instance ou personnalité politique ou syndicale) et/ou parties successives dans un corpus homogène (par ex. dynamique interne du texte, de son début à sa fin), etc.

[86] Jacques Jenny, o.c., mars 1997

[87] Jacques Jenny, o.c., mars 1997

[88] Jacques Jenny, o.c., mars 1997

[89] La Libre Belgique, 3 mai 2007.

communiqué	2	intervenue	1
Copiepresse	2	jeudi	1
des	2	la	1
deux	2	membres	1
et	2	moteur	1
Google	2	ont	1
les	2	par	1
un	2	parties	1
accord	1	précisent	1
annoncé	1	presse	1
belge	1	prise	1
cadre	1	qui	1
cette	1	quotidienne	1
constructif	1	recherche	1
d'	1	référencement	1
décision	1	repris	1
dialogue	1	reprise	1
éditeurs	1	sites	1
entre	1	sociétés	1
entreprises	1	sur	1
est	1		

Il s'agit d'une démarche mathématique « élémentaire » et de type lexical.

Elle peut cependant faire intervenir des aspects syntaxiques et sémantiques afin de lever les ambiguïtés du vocabulaire.
La lexicométrie est une approche bottom-up. Ce sont les traitements des données qui guident l'interprétation et la production de savoirs.[90]

Différents problèmes peuvent se poser : savoir ce qu'on compte (mots, expressions, mots composés), l'identification des formes lexicales, le regroupement de certaines formes, l'utilisation d'un dictionnaire ou d'un thésaurus, l'homonymie (désambiguïsation), etc.

L'exemple présenté ci-dessus le montre bien. Certaines formes devraient sans doute être regroupées ensemble tels « a » et « ont » sous la forme plus générale « avoir ».

De plus, on peut se poser la question de la signification des résultats obtenus et de leur interprétation.[91]

Dans notre exemple, les termes les plus fréquents ne représentent pas des thèmes. Il s'agit d'articles « le » et « de ». Il convient donc de faire un tri dans les résultats.

Enfin, la lexicométrie ne peut répondre qu'à la question du « quoi » (que dit-il ?) et non aux questions du « comment » et du « pourquoi ».[92]

1.2. L'analyse statistique des données textuelles

A côté de la lexicométrie classique, diverses méthodes statistiques applicables à des matériaux textuels sont apparues.

[90] R. Ghiglione et al., o.c., p. 6.
[91] R. Ghiglione et al., o.c., p. 3.
[92] R. Ghiglione et al., o.c., p. 5.

L'application d'outils statistiques à l'analyse de texte permet de confirmer ou d'approfondir les résultats obtenus via la lexicométrie décrite ci-dessus. Elle constitue donc un prolongement de celle-ci

Elle comprend notamment les méthodes suivantes.

L'*analyse factorielle* est une « famille de méthodes statistiques d'analyse multidimensionnelle, s'appliquant à des tableaux de nombres, qui visent à extraire des "facteurs" résumant approximativement par quelques séries de nombres l'ensemble des informations contenues dans le tableau de départ ». [93]

L'*analyse factorielle des correspondances* (AFC) est une « méthode d'analyse factorielle s'appliquant à l'étude de tableaux à double entrée composés de nombres positifs. L'AC est caractérisée par l'emploi d'une distance (ou métrique) particulière dite distance du chi-2 (ou c2) ». [94]

« Sa finalité est de trouver le meilleur résumé possible dans un espace de dimensions réduites. Ce meilleur résumé est constitué d'un petit nombre d'axes qui maximise l'inertie projetée Elle va permettre une représentation simultanée des lignes et des colonnes dans l'espace de dimensions réduites cherché. »[95]

La *Classification Ascendante Hiérarchique* (CAH) est une « méthode de classification hiérarchique partant des individus isolés assimilés à des classes et procédant, à chaque étape, par agrégation des deux classes les plus proches au sens de la norme choisie. Chaque niveau de hiérarchie représente une classe. Un arbre planaire hiérarchique permet de décrire de façon explicite la structure finale de la classification obtenue : "plus les individus se regroupent bas dans l'arbre, plus ils se ressemblent" ». [96]

« Le principe des algorithmes de classification hiérarchique ascendante est très simple :
Initialisation : chaque élément de constitue une classe. Une "distance" D est calculée entre toutes les classes.
Tant que nombre de classes > 1
– regrouper les deux classes les plus proches au sens de la "distance" D,
– calcul des "distances" entre la nouvelle classe et les autres. »[97]

La *classification descendante hiérarchique* (CDH) « procède par fractionnements successifs du texte. Elle repère les oppositions les plus fortes entre les mots du texte et extrait ensuite des classes d'énoncés représentatifs.[98] Celle-ci aboutit à une série de classes construites de manière formelle. »[99]

[93] B. Fracchiolla, A. Kuncova, A. Maisondieu, o.c.
[94] B. Fracchiolla, A. Kuncova, A. Maisondieu, o.c.
[95] http://www.obs-vlfr.fr/Enseignement/enseignants/labat/anado/afc/presentation.html
[96] http://mist.univ-paris1.fr/logiciel/def.htm#motassoc1
[97] C. Ambroise, « Classification hiérarchique », http://www.hds.utc.fr/sy09/documents/hierarchie.pdf
[98] On part de la totalité du texte, et on découpe ce texte en unités textuelles, ces unités représentent des morceaux de texte dont la taille est d'ordre de la phrase. A partir de ces unités textuelles, on va ensuite dissocier deux groupes d'unités dont les vocabulaires sont les plus différents possibles. Ces deux groupes obtenus en utilisant la métrique du khi2. Alceste repère ensuite le plus grand des deux groupes et continue le processus, de manière itérative, jusqu'à l'obtention d'un nombre de classes généralement prédéfinies à l'avance. http://www.image.cict.fr/index_alceste.htm
[99] http://www.image.cict.fr/index_alceste.htm

L'*analyse en composantes principales* (ACP) est « une technique mathématique permettant de réduire un système complexe de corrélations en un plus petit nombre de dimensions ».[100]

« A partir d'un ensemble n d'objets dans un espace de p descripteurs, son but est de trouver une représentation dans un espace réduit de k dimensions (k << p) qui conserve "le meilleur résumé " (au sens du maximum de la variance projetée). »[101]

« Le but de l'analyse en composantes principales est de réorganiser les données de telle manière qu'elles ne soient plus corrélées (c.-à-d. qu'elles deviennent indépendantes). »[102]
« En ne conservant que les composantes les plus significatives pour l'analyse, il est possible de réduire considérablement le volume de données à traiter. » [103]

La *classification par partition* (CPP) « décompose l'ensemble en un nombre de classes fixé à priori, initialisé par un ou plusieurs représentants. Elle utilise des algorithmes de classification non hiérarchique (le calcul de l'inertie interclasse et intraclasse, le regroupement autour des centres mobiles et la méthode des nuées dynamiques). Elle est basée sur l'existence d'un critère global qui mesure la distance entre les individus et par la même, la qualité d'une partition. Les résultats sont présentés sous forme de cartes factorielles. »[104]

2. L'analyse socio-sémantique

La deuxième catégorie de méthodologie d'analyse de texte est l'analyse socio-sémantique.

Dans l'analyse socio-sémantique, une grille d'analyse des contenus thématiques du corpus, en fonction du cadre théorique de la recherche, est établie préalablement sur base d'un savoir antérieur.
Les traitements effectués ultérieurement, statistiques ou non, ne sont là que pour mesurer l'« extension empirique » des catégories établies a priori, puis pour valider, ou non, les hypothèses préalables du chercheur.
L'analyse est donc établie sur des savoirs préalables qui serviront à étayer l'interprétation. [105]

Étant donné la multitude de savoirs et le fait que chaque recherche se base sur un savoir plus ou moins différent, il est sans doute plus opportun de parler de cette méthode au pluriel.

Sur base de ce savoir préalable, le chercheur va établir une grille d'analyse des contenus thématiques qu'il s'attend à retrouver dans le corpus étudié. Les thèmes sont donc préexistants et l'analyse permettra de confirmer, infirmer ou compléter ceux-ci.

Exemple : si on analyse des interviews portant sur la vie professionnelle, on peut s'attendre à trouver certains thèmes : relations de travail avec les supérieurs, les collègues, les subordonnés, les conditions de travail, etc.

Contrairement à la lexicométrie, il s'agit d'une approche top-down. Ce sont les savoirs *a priori* qui guident le traitement des données. [106]

[100] http://www.ulg.ac.be/pedaexpe/cours/glosaire/acp.htm
[101] http://www.obs-vlfr.fr/Enseignement/enseignants/labat/anado/acp/presentation.html
[102] http://telsat.belspo.be/beo/fr/guide/compprin.asp?section=3.10
[103] http://telsat.belspo.be/beo/fr/guide/compprin.asp?section=3.10
[104] http://mist.univ-paris1.fr/logiciel/def.htm
[105] R. Ghiglione et al., o.c., p. 6.
[106] R. Ghiglione et al., o.c., p. 6.

Au niveau méthodologique, elle opère « *par segmentation du corpus en unités de signification pertinentes et par catégorisation multidimensionnelle conforme aux grilles d'analyse conceptuelle spécifiques de chaque recherche (dans une optique classique de codage a posteriori, où le chercheur lit le texte, "marque" et code lui-même les unités de sens du corpus), et par recours éventuel à des méthodes statistiques, notamment d'inspiration booléenne comme les "arbres de décision" ou le "data mining" à la mode, plus diversifiées que celles de la seule méthode benzécriste* ».[107]

Pratiquement, le corpus est découpé en blocs, et chaque bloc est assorti d'un ou plusieurs thèmes identifiés préalablement dans la grille. Éventuellement, des calculs statistiques sont appliqués aux résultats.

Cette méthode laisse certaines questions en suspens. Comment construire les grilles d'analyse préalables ?, Sur quelle théorie se baser pour réaliser les interprétations ?, etc.[108]

3. L'analyse par réseau de mots associés (co-word analysis)

L'analyse par réseau de mots associés est une « méthode basée sur le calcul des fréquences de cooccurrence des termes du corpus, afin de mettre en évidence la structure de leurs relations (clusters). Elle se propose d'identifier les mots les plus fortement associés entre eux, conduisant à des termes de recherche, donc à une classification des contenus. » [109]

Son objectif est d'identifier les thèmes qui sont en relation dans le texte, et donc dans l'esprit de son auteur. Elle peut également servir à mettre en évidence les dissociations ou exclusions de thèmes.

Elle correspond à l'analyse des cooccurrences définie dans le premier chapitre.

Il s'agit d'une méthode quantitative puisque certains décomptes sont opérés.
Les associations ne porteront toutefois pas nécessairement sur les termes les plus fréquents. Le comptage portera plus sur le nombre de fois que deux termes sont associés que sur le nombre absolu de chaque terme dans le corpus.

La méthode est la suivante.

« L'association de deux mots-clés se mesure en fonction de leur nombre d'apparitions communes dans les documents qu'ils indexent.
Les associations sont affectées d'une valeur et tous les couples de termes obtenus sont triés par valeurs décroissantes. Les clusters sont construits à partir de la liste classée des couples.
L'ensemble des éléments à agréger forment initialement un seul grand réseau d'association. Il s'agit d'un réseau valué c'est-à-dire un système dans lequel les mots sont reliés par des liens plus ou moins forts.

- Si un couple de termes appartient au même cluster, le lien entre ces termes est considéré comme un lien interne au cluster.
- Si les termes d'un couple appartiennent à deux clusters différents, leur lien est considéré comme lien externe, lien entre clusters.

A la suite de la classification des mots-clés, les documents sont affectés aux clusters.

---

[107] Jacques Jenny, o.c., Mars 1997
[108] R. Ghiglione et al., o.c., p. 6.
[109] http://mist.univ-paris1.fr/logiciel/def.htm#motassoc1

Les clusters sont situés dans un espace bi-dimensionnel et dans un plan défini par un coefficient de cohérence interne c'est-à-dire de *densité* (moyenne des valeurs des associations "internes". Plus elle est élevée, plus le cluster est considéré comme étant un agencement bien structuré et reconnu) et par un coefficient de *centralité* (qui exprime la capacité de connexion, la puissance de capture, d'affecter ou d'être affecté du même cluster). »[110]

Cette méthodologie est de nature lexicale et de type bottom-up.[111]

Certaines critiques peuvent être formulées à son égard.

L'on peut se demander à partir de quand deux termes sont associés.
Pour repérer ces associations, on peut procéder de deux manières. Soit on découpe préalablement le corpus en « blocs », et deux termes sont cooccurrents s'ils sont dans le même « bloc ». Soit, on part d'un terme et l'on cherche dans son « contexte » (x mots avant et x mots après) quels autres termes lui sont associés.
Dans la première méthode, le résultat sera dépendant du découpage de départ et certaines associations pourront être manquées. Dans la deuxième, il convient de déterminer la taille du « contexte ». Trop grand, il n'aura plus de sens ; trop petit, l'on risque de passer à côté de certaines choses.

4. L'analyse cognitivo-discursive

L'analyse cognitivo-discursive est une méthode d'analyse de contenu de type qualitative à vocation principalement thématique.

Partant de l'idée que toute production langagière (orale ou écrite) est une communication, l'analyse cognitivo-discursive a pour objectif d'identifier le projet de sens et l'intention présents dans ladite communication.[112] Il s'agit donc d'identifier de quoi parle le texte, et comment l'auteur en parle, mais également le ou les buts poursuivis par ce dernier. Cette méthode veut donc aller beaucoup plus loin que les méthodes précédentes qui ne visent généralement qu'à identifier des thèmes, et donc à répondre aux questions du qui et du quoi.

Elle est issue de la synthèse de deux autres méthodes qualitatives que nous allons présenter en premier lieu : l'analyse propositionnelle du discours (APD) et l'analyse propositionnelle prédicative (APP).

Tant l'APD que l'APP visent à décrire « les logiques de construction progressive de tout univers référentiel cohérent, ainsi que les finalités ou intentions de chaque mise en scène langagière particulière ».[113]

Avant cette présentation, quelques remarques préalables s'imposent.
Les trois méthodes que nous allons décrire ici sont qualitatives. Elles n'utilisent pas de décompte ou de calcul statistique pour produire des résultats, contrairement à la plupart des méthodes décrites ci-avant. Elles visent plutôt à attribuer des caractéristiques qualitatives à certains éléments du texte.

[110] http://mist.univ-paris1.fr/logiciel/def.htm#motassoc1
[111] R. Ghiglione et al., o.c., p. 8.
[112] R. Ghiglione et al., o.c., p. 14.
[113] R. Ghiglione et al., o.c., p. 2.

Ensuite, elles ont pour unité d'analyse la proposition. Celle-ci est définie comme un sujet plus un verbe et un ou deux compléments. La proposition se distingue donc de la phrase (celle-ci pouvant être composée de plusieurs propositions). Par ce point, elles se distinguent donc aussi des méthodes précédentes qui ont pour unité d'analyse le mot.

4.1. L'analyse propositionnelle du discours (APD)[114]

L'analyse propositionnelle du discours est présentée par Laurence Bardin comme « *une variante de l'analyse thématique, cherchant à dépasser certaines insuffisances du découpage en catégories* ».[115]

Selon Ghiglione, l'objectif de l'APD est de répondre à la question : « Comment un sujet traite-t-il l'information, qu'il la reçoive ou qu'il la produise ? ».[116] Elle veut répondre aux questions : qui, quoi, comment, pourquoi.

La réponse apportée par Ghiglione est la suivante : « *Le sujet traite l'information en mettant en scène un ensemble structuré et plus ou moins cohérent de micro-univers, chacun étant constitué d'une scène peuplée a minima d'un actant qui fait l'action (placé le plus souvent en position de sujet) et de l'acte que le verbe accomplit* ».[117]

L'aspect « mise en scène » est important. Lorsqu'une personne s'exprime, que ce soit oralement ou par écrit, elle met en scène une ou plusieurs représentations qu'elle a d'un sujet donné. Cette mise en scène traduit une certaine logique qu'à l'auteur. L'APD s'intéresse donc à la logique de l'acteur.

L'idée sous-jacente est que le langage porte des indicateurs de ces représentations.

La méthode consiste donc à atteindre ces représentations et leur mise en scène en analysant les propositions du texte et les différents éléments qui les composent.

Ces représentations peuvent être identifiées par le repérage de micro-univers.

Ces micro-univers correspondent à l'identification de « référents-noyaux », c'est-à-dire de substantifs ou pronoms constituant des « pôles d'attraction sémantique structurant l'ensemble des paroles dans un contexte donné ». Ces référents-noyaux sont vus en terme d'« acteurs ».
Concrètement, il s'agit de repérer les substantifs et pronoms occupant une place centrale dans le texte, en terme qualitatif et non quantitatif. Tous ne sont donc pas retenus.
Cela permet d'identifier les thèmes abordés par l'auteur du texte. Les relations entre ceux-ci sont aussi analysées, ainsi que leur position d'actant ou d'acté (avant ou après un verbe). [118]
Cette première étape permet de répondre à la question du « qui ».

L'APD code également les autres éléments de la proposition afin d'atteindre les représentations sous-jacentes au texte.[119]

---

[114] Un exemple figure à l'annexe 2.
[115] L. Bardin, o.c., p. 243.
[116] R. Ghiglione et al., o.c., p. 65.
[117] R. Ghiglione et al., o.c., p. 65.
[118] Jacques Jenny, o.c., mars 1997
[119] R. Ghiglione et al., L'analyse automatique des contenus, Paris, Dunod, 1998, p. 66-67.

L'analyse des verbes permet de s'intéresser aux actes que les acteurs (référents-noyaux identifiés ci-dessus) accomplissent. Ils sont codés selon leur temps, leur fonction (factif, statif, déclaratif, etc.) et leur polarité (positive ou négative).
Cette analyse permet de répondre à la question du « quoi ».

Les adjectifs permettent d'exprimer les caractéristiques des acteurs. Combinés aux substantifs et verbes, ils permettent de compléter les réponses aux « qui » et « quoi ».

Enfin, l'analyse des joncteurs (conséquence, causalité, but, addition, disjonction, etc.) et modalisations (adverbes) permet de répondre à la question du « comment ». [120]

L'APD analyse donc chaque terme, mais dans le cadre de son contexte propositionnel, et non de manière indépendante comme dans la plupart des autres méthodes vues ci-avant.

Dans le cadre de cette méthode, l'unité d'analyse est donc la proposition grammaticale. Les analyse effectuées ci-dessus peuvent être complétés par une sélection des phrases clés du corpus.
Une proposition est une phrase qui qualifie, explique un référent-noyau. [121] Ainsi, sont sélectionnées les propositions se rattachant aux référents-noyaux identifiés. Elles sont par la suite réécrites sous forme simplifiée ; puis leur nombre est réduit par « éliminations justifiées » (phrases synonymes, propositions emboîtées, etc.). [122]
Les propositions restantes correspondent au noyau du texte.

L'APD met donc en jeu des analyses lexicales, mais également syntaxiques et sémantiques.

Une critique qui peut-être formulée est qu'il ne semble pas y avoir de critères précis afin de déterminer les substantifs et pronoms centraux dans un corpus. Est-ce leur position de sujet ou de complément dans la proposition qui va être déterminante ? Nous n'avons pas trouvé de réponse à cette question ?

4.2. L'analyse prédicative du discours (APP)

L'analyse prédicative du discours « *porte sur l'inscription des propositions de forme prédicat/argument, et sur la hiérarchisation du texte à partir de macro-propositions* ». [123]

« *… elle traduit le texte en relations prédicatives du genre "que dit-on à propos de quoi ?" - ce qui correspond aux relations entre le "Thème" (de quoi ça parle ?) et le "Rhème" (ce qu'on en dit), constitutives de tout discours*, (…) ». [124]

Une proposition prédicative est une construction de la forme : prédicat et un ou plusieurs arguments.[125] Un prédicat peut être un verbe, mais également un adjectif, un attribut, un connecteur, une préposition, etc. Les arguments sont généralement des substantifs ou d'autres propositions.

Exemple : « L'oiseau chante » : CHANTER (oiseau)

---

[120] L. Bardin, o.c., p. 250.
[121] L. Bardin, o.c., p. 244.
[122] L. Bardin, o.c., p. 245.
[123] Jacques Jenny, o.c., mars 1997
[124] Jacques Jenny, o.c., mars 1997
[125] R. Ghiglione et al., o.c., p. 24.

« Socrate est un philosophe » : PHILOSOPHE (Socrate)

L'APP commence donc par découper le texte en propositions sous forme prédicative.

Il existe différentes formes de propositions prédicatives ayant des statuts différents :
- Les propositions où le prédicat est un verbe
  Exemple : « Le président a été élu par les électeurs » : ELIRE (président, électeurs)
- Les propositions où le prédicat est un connecteur
  Exemple : « Un accident s'est produit car il pleuvait et la chaussée était dégradée »
  (1) SE PRODUIRE (accident)
  (2) CAR ((3),(4))
  (3) PLEUVOIR ()
  (4) ETRE DEGRADE (chaussée)
- Les propositions où le prédicat est un adjectif, une préposition ou un adverbe
  Exemple : « Je suis venue ce soir »
  (1) VENIR (je)
  (2) CE SOIR (1)

L'APP considère qu'il y a une hiérarchie entre ces différentes formes, les deux premières étant plus importantes, centrales, et la dernière secondaire, périphérique.

Donc, elle « *procède (…) à la hiérarchisation des propositions, depuis la proposition topique, de niveau zéro - qui est perçue comme la plus importante - jusqu'à celles de niveaux subséquents qui conservent au moins un argument commun avec celles du niveau précédent.* »[126]

Comme l'APD, l'APP met en jeu des analyses lexicales, mais également syntaxiques et sémantiques. Mais contrairement à l'APD, l'APP s'intéresse à la logique du texte. [127]

4.3. L'analyse cognitivo-discursive (ACD)

L'analyse cognitivo-discursive constitue la synthèse de l'APD et de l'APP.

Comme cela a déjà été dit, elle vise à retrouver « le projet de sens et l'intention inscrits dans un contrat de communication spécifique ».[128]

Héritière de l'APD, elle « décrit les logiques de construction progressive de tout univers référentiel cohérent, avec la notion de "schéma causal", ainsi que les finalités ou intentions de chaque mise en scène langagière particulière, avec différents "opérateurs argumentatifs" ».[129]
Elle vise à atteindre la « structure fondamentale de la signification », c'est-à-dire le sens et l'intention présents dans le texte.

Héritière de l'APP, elle procède à une hiérarchisation des propositions du texte sur base de leur position centrale ou périphérique.
Les propositions prédicatives avec verbe ou connecteur posent les événements principaux ouvrant et clôturant l'action. Les propositions prédicatives avec adjectif, préposition ou

[126] Jacques Jenny, o.c., mars 1997
[127] R. Ghiglione et al., o.c., p. 65.
[128] R. Ghiglione et al., o.c., p. 14.
[129] Jacques Jenny, o.c., mars 1997

adverbe, au contraire, anecdotisent le propos.[130] Triées selon cette hiérarchie, seules les propositions centrales sont retenues pour atteindre le « noyau générateur de la référence ».

La combinaison des deux, « structure fondamentale de la signification » et « noyau générateur de la référence », constitue le principe de l'ACD, permettant d'atteindre le sens et le but d'un texte.

Selon Ghiglione, l'ACD réconcilie les approches bottom-up et top-down. En effet, son application conduit à extraire des éléments dans une perspective bottom-up, éléments qui « *permettent* de réinscrire un discours particulier dans une interdiscursivité qui permet, quant à elle, des interprétations sécurisées par un traitement *top-down »* .[131]

Quant à l'origine du terme ACD, il provient de l'intérêt de ses concepteurs pour « les structures et moyens **cognitifs** mis en place par le sujet psycho-social au fil de ses divers apprentissages et inscriptions, pour produire et traiter du **discours**, par nature interlocutoire ».[132]

5. Synthèse

Le présent point va faire une synthèse des différentes méthodes présentées ci-dessus. Gardons à l'esprit qu'il s'agit de méthodes pour lesquelles des logiciels informatiques ont pu être identifiés.

Il est possible de classer ces méthodes selon différents critères.

Le premier de ces critères est le type d'approche suivi : quantitatif ou qualitatif.

Dans l'approche quantitative, on se base sur la « fréquence d'apparition de certaines caractéristiques de contenu ».[133]
Dans l'approche qualitative, on prend en considération « la présence ou l'absence d'une caractéristique de contenu donnée ou d'un ensemble de caractéristiques, dans un certain fragment de message ».[134]

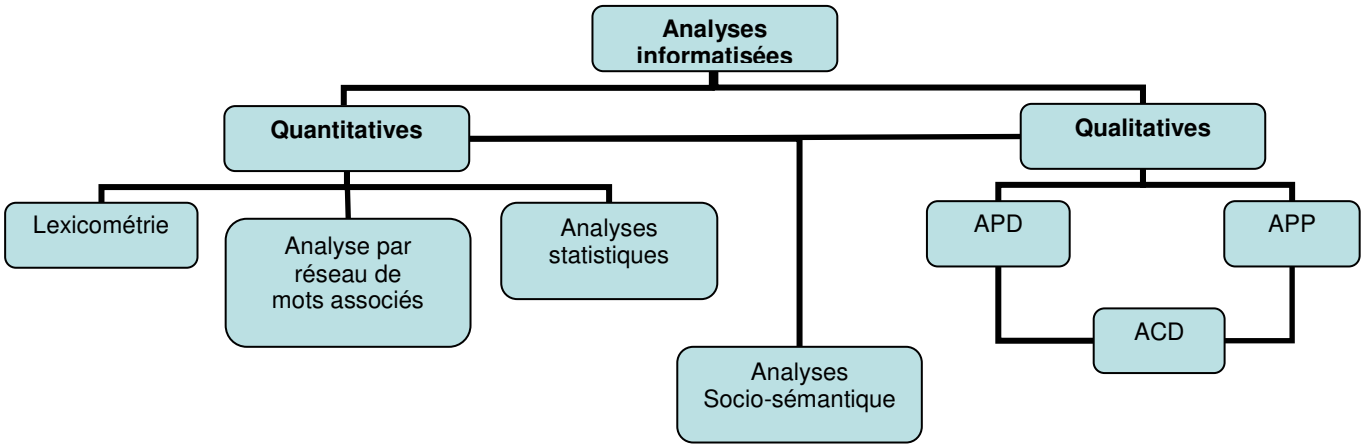
Le débat entre quantitatif et qualitatif n'est pas nouveau et persiste de nos jours.[135]

Le schéma ci-dessous donne un aperçu de la répartition des méthodes informatisées selon ce critère. Ainsi, la lexicométrie, les analyses statistiques et les analyses par réseau de mots associés sont clairement quantitatives étant donné qu'elles se basent sur des décomptes et calculs. Les analyses propositionnelle du discours, propositionnelle prédicative et cognitivo-discursive sont quantitatives car elles attribuent des caractéristiques non chiffrables aux éléments du texte (analyse des termes, hiérarchisation des propositions).

Toutefois, certaines méthodes ont des aspects tant quantitatifs que qualitatifs. L'analyse socio-sémantique se base sur une grille d'analyse préalable et vise à appliquer cette grille aux

[130] R. Ghiglione et al., o.c., p. 73.
[131] R. Ghiglione et al., o.c., p. 9.
[132] R. Ghiglione et al., o.c., p. 14.
[133] L. Bardin, o.c., p. 24.
[134] L. Bardin, o.c., p. 24.
[135] L. Bardin, o.c., p. 24.

éléments du texte (les blocs en l'occurrence), donc à leur attribuer une qualité, en l'espèce, un des éléments de la grille. Elle peut toutefois être complétée par des analyses quantitatives.



Il convient cependant de mentionner que l'opposition entre « qualitatif » et « quantitatif » est remise en cause par Jenny.

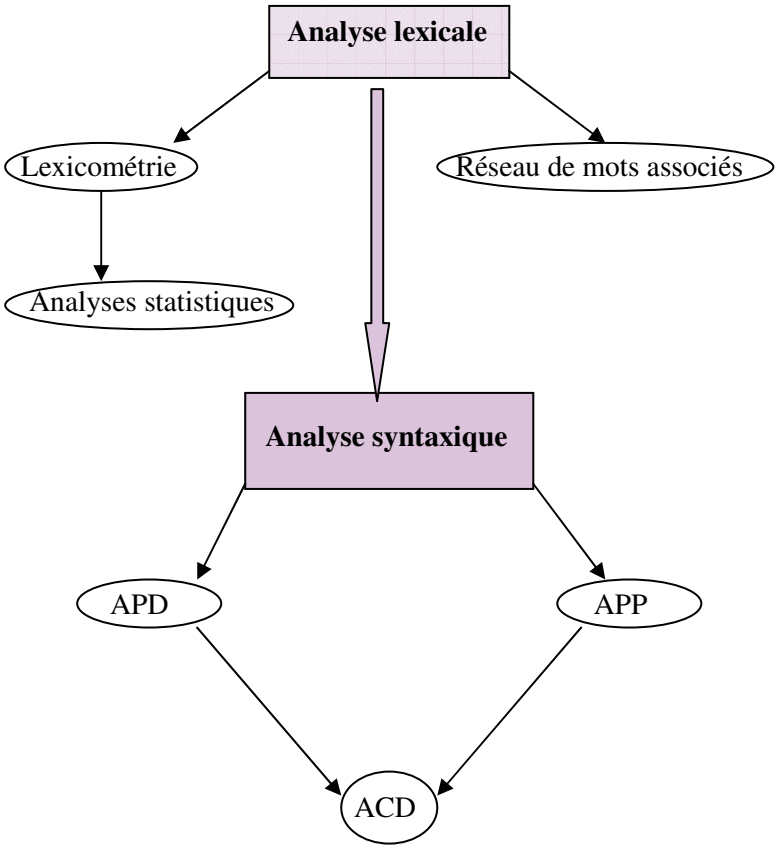
« *Tout d'abord, nous considérons que* ***toute*** *recherche sociologique comporte nécessairement une part de "****matériaux textuels****" à analyser - au point que la distinction entre le "qualitatif" et le "quantitatif" ne saurait être au mieux qu'une distinction de phases, de moments dans la recherche, et au pire qu'une mystification destinée peut-être à masquer les méconnaissances respectives de la réalité discursive (ou de la dimension "intensive", pour reprendre l'expression de l'épistémologue Canguilhem (1950), à propos de l'analyse conceptuelle) chez les quantitativistes et de la réalité numérique (ou de la dimension "extensive", avec les incontournables "opérateurs de quantification") chez les qualitativistes.* »[136]

« *Ce choix exprime le refus de considérer les méthodes dites quantitatives et qualitatives comme des méthodes alternatives, voire opposées, et la conviction qu'il s'agit de catégories du sens commun, superficielles et fallacieuses; avec la volonté de promouvoir une conception méthodologique fondée sur la synthèse, l'interpénétration, la fécondation mutuelle, de ces deux* ***modes d'expression complémentaires*** *d'une seule et même "réalité sociale".* » [137]

Un autre critère de classification est l'unité d'analyse choisie : le mot ou la phrase, et donc l'analyse lexicale ou syntaxique, étant entendu qu'il paraît difficile de faire de l'analyse syntaxique sans faire de l'analyse lexicale et que l'analyse lexicale devra parfois faire le détour par l'analyse syntaxique pour lever les ambiguïtés.
Il convient également de citer l'analyse sémantique, la recherche du sens. Il apparaît évident que ces différentes méthodes ne peuvent faire l'impasse sur celle-ci (pour lever des ambiguïtés qu'une analyse syntaxique n'aurait pu faire, par exemple).

Le schéma ci-dessous donne un aperçu de la répartition en fonction de ce critère.

[136] Jacques Jenny, o.c., Mars 1997
[137] Jacques Jenny, o.c., Mars 1997

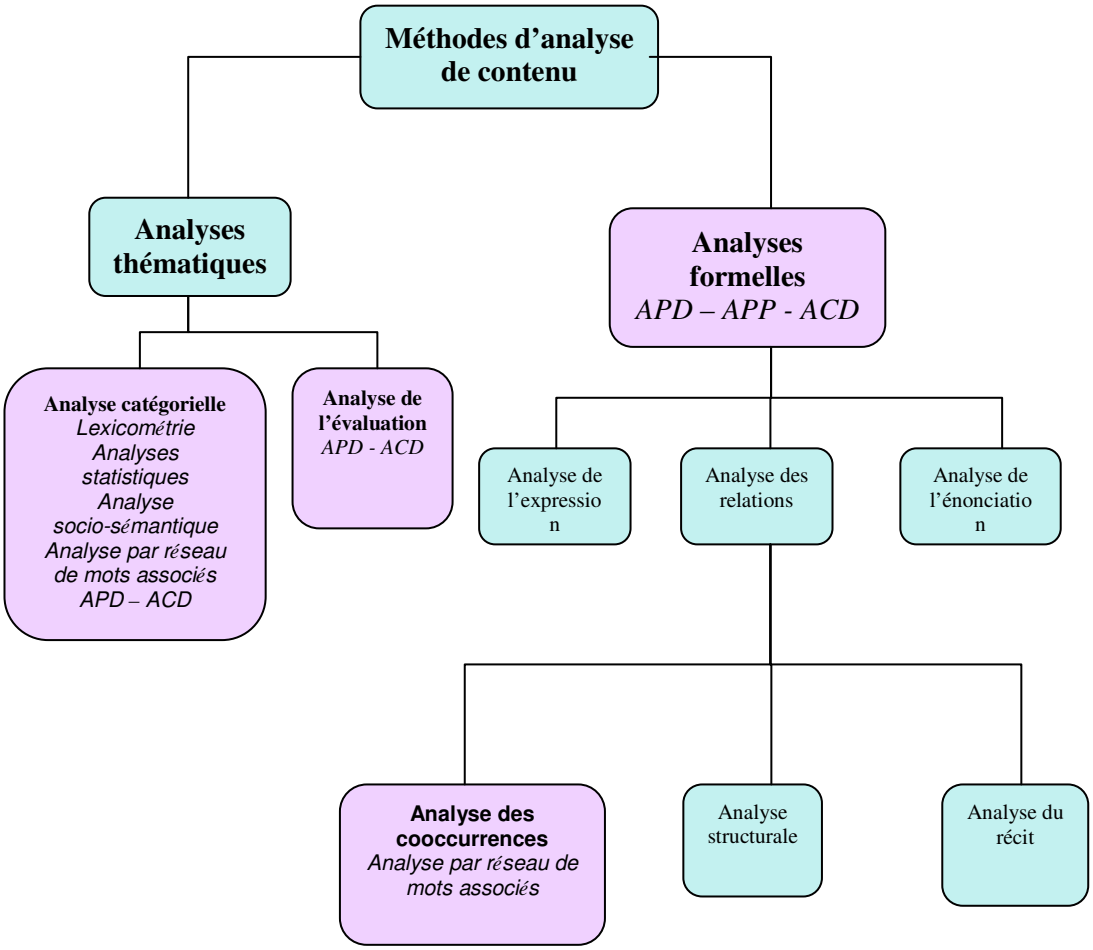


La lexicométrie, les analyses statistiques et l'analyse par réseau de mots associés ont pour unité d'analyse le mot. Celui-ci est considéré indépendamment des propositions dans lesquelles il est contenu. Il est analysé pour lui-même (quant à sa présence quantitative ou par rapport aux relations qu'il entretient avec d'autres mots).

Les analyseS propositionnelle du discours, propositionnelle prédicative et cognitivo-discursive ont pour unité d'analyse la proposition. Et si une analyse des mots se fait dans le cadre de l'APD et l'ACD, c'est en relation avec la place de ceux-ci dans la proposition (actant, acté) ou avec le rôle qu'ils peuvent jouer en reliant plusieurs de celles-ci (connecteurs).

Quant à l'analyse socio-sémantique, elle se classe encore à part puisque son unité d'analyse est le bloc de texte. Elle mettra toutefois en jeu des analyses lexicales, syntaxiques et sémantiques selon le savoir préalable qui la guide.

Enfin, nous allons classer les méthodes informatisées au sein des méthodes d'analyse de contenu.



Pour rappel, les types d'analyse de contenu que nous avons décrits dans le premier chapitre se regroupent en deux grandes catégories. D'un côté, les analyses thématiques se penchent sur le fond du texte pour en découvrir l'objet et la manière dont l'auteur en parle. De l'autre, les analyses formelles étudient la structure d'un texte afin d'en repérer les différents éléments ainsi que leur agencement, ceux-là devant éclairer sur l'état d'esprit de l'auteur lors de la rédaction du texte.

Si nous reprenons les méthodes informatisées étudiées dans ce deuxième chapitre, nous remarquons que la plupart ont pour objectif d'identifier les thèmes abordés par un texte, d'une manière ou d'une autre. Ainsi, elles peuvent être rattachées à l'analyse catégorielle, sous-catégorie de l'analyse thématique.
La lexicométrie et les analyses statistiques textuelles prétendent identifier ces thèmes de manière quantitative, sur base de décomptes et de calculs.

L'analyse par réseau de mots associés permet de compléter les approches précédentes en découvrant les relations pouvant exister entre thèmes, à nouveau sur base de décomptes.
Dans le cadre de l'analyse socio-sémantique, une grille préalable des thèmes possibles d'un corpus est établie sur base des savoirs présidant la recherche. Le codage du corpus selon cette grille permettra de la confirmer, de la modifier, voire de l'infirmer.
Enfin, les analyses propositionnelle du discours et cognitivo-discursive ont pour objet de mettre à jour les représentations de l'auteur sur un sujet donné, et donc, les thèmes abordés.

Par contre, nous estimons que l'analyse propositionnelle prédicative n'en fait pas partie car son analyse porte principalement sur la hiérarchisation des propositions du corpus et donc sur la forme du texte, plutôt que sur le fond.

Par ailleurs, les analyses propositionnelle du discours et cognitivo-discursive peuvent également être classées parmi l'analyse de l'évaluation dans la mesure où elles tendent également à identifier la manière dont l'auteur parle des thèmes, objets du texte.

L'analyse par réseau de mots associés se rattache également à l'analyse des cooccurrences. Elle en constitue même la traduction méthodologique. Elle met en avant les relations de proximité entre thèmes.

Enfin, les analyses propositionnelle du discours, propositionnelle prédicative et cognitivo-discursive se rattachent également aux analyses formelles. En effet, elles analysent certains éléments structurels de la phrase (liens entre phrases via les connections ou joncteurs, place des éléments dans la proposition, etc.) et leurs relations.

Après avoir posé cet état de l'art en matière d'analyse de texte, nous allons, dans la deuxième partie de ce mémoire, nous tourner vers le terrain en nous plongeant au cœur des logiciels d'analyse de texte, mais également en nous tournant vers les utilisateurs potentiels de tels logiciels.

Partie 2 : Les logiciels d'analyse de texte et les attentes des chercheurs

La première partie de ce mémoire nous a permis de poser les bases théoriques et de comprendre la notion d'analyse de texte.

La deuxième partie va se pencher sur le terrain en permettant de se familiariser avec les logiciels d'analyse de texte, mais également en découvrant les attentes du monde de la recherche à leur égard.

Cette deuxième partie sera divisée en trois chapitres.

Le chapitre 3 aura pour objectif de présenter les logiciels d'analyse de texte existants actuellement. Ces logiciels seront classés selon le(s) type(s) de méthode d'analyse qu'ils implémentent. La liste étant longue, une description de chacun d'eux figure en annexe. Cependant, quelques-uns seront sélectionnés et analysés en profondeur selon des critères préalablement établis. Ils seront également comparés et évalués.

Le chapitre 4 portera sur la découverte des attentes des chercheurs par rapport à ces logiciels. Différentes interviews ont été menées et analysées. Les résultats de ces analyses seront présentés dans le cadre de ce chapitre.

Enfin, le chapitre 5 tentera de répondre à la question de savoir si les logiciels du chapitre 3 peuvent répondre aux attentes cernées dans le chapitre 4.

Chapitre 3 : Présentation de quelques logiciels d'analyse de texte

L'objectif de ce troisième chapitre est de donner un aperçu des logiciels d'analyse de texte existants. Il s'agira de se familiariser avec les outils que ces logiciels offrent, de cerner leurs potentialités et leurs limites. Il s'agira également de percevoir la variété des fonctionnalités proposées.

L'identification de ces logiciels repose principalement sur des listes établies par :

- Christophe Lejeune[138] ;
- Le département « Maîtrise des Sciences de l'information et de la documentation » de l'Université Paris 1 – Panthéon-Sorbonne[139] ;
- Cyril Gruau[140] ;
- Jacques Jenny.[141]

Ce travail d'identification s'est complété d'une collecte de renseignements sur chaque logiciel : site Internet dédié, manuels, articles en ligne, livres, etc., afin d'en cerner toutes les caractéristiques.

Cyril Gruau[142] distingue les logiciels d'analyse quantitative et ceux d'analyse qualitative. Ce sont les derniers qui nous intéressent dans le cadre de ce mémoire.

Paris ceux-ci, il identifie deux grandes familles :

- Les logiciels issus du monde francophone et qui, bien que de type qualitatif, se basent sur les statistiques (dénombrement, classification) pour opérer leurs analyses ;
- Les logiciels issus des pays-anglo-saxons, les CAQDAS (**C**omputer-**A**ssisted **Q**ualitative **D**ata **A**nalysis **S**oftware), qui procèdent par étiquetage et requêtage. « Un corpus de texte subit un codage (parfois partiellement automatisé) qui consiste à identifier des mots, expressions, phrases relevant d'un thème, d'une figure, d'un sujet ou d'une caractéristique linguistique. Les codes servent ensuite à repérer des associations régulières ou originales et sont éventuellement mobilisés dans des applications quantitatives. »[143]

Dans le cadre de ce mémoire, nous aborderons principalement, mais non exclusivement, les logiciels du monde francophone.

La description des logiciels retenus figure en annexe (Annexe 3) par ordre alphabétique. Pour des raisons évidentes, seuls les logiciels gratuits ou disposant d'une version de démonstration ou d'évaluation ont été retenus.

Pour chacun d'eux, nous avons essayé, dans la mesure de la disponibilité des ressources[144], de présenter les principales caractéristiques (langues, système d'exploitations, etc.), ainsi que les domaines d'application et les fonctionnalités offertes.

---

[138] http://analyses.ishs.ulg.ac.be/logiciels/
[139] http://mist.univ-paris1.fr/logiciel/frame.htm
[140] http://www.cemef.net/fr/presentation/pagesperso/cg-promo2001/extdoc/Gruau-AnalyseQualitative.pdf
[141] http://pageperso.aol.fr/jacquesjenny/ATBMS.htm ; http://pageperso.aol.fr/jacquesjenny/ClassifLogicielsAT06.htm
[142] http://www.cemef.net/fr/presentation/pagesperso/cg-promo2001/extdoc/Gruau-AnalyseQualitative.pdf
[143] C. Lejeune, http://analyses.ishs.ulg.ac.be/logiciels/index_2.html
[144] Certains logiciels commerciaux ne fournissent pas de manuel gratuit en ligne, par exemple.

En ce qui concerne la suite de ce chapitre, un premier point présentera la classification de ces logiciels selon les méthodes informatisées identifiées dans le chapitre deux.
Un deuxième point aura pour objectif de décrire de manière détaillée les fonctionnalités offertes par quelques-uns de ces logiciels.

1. Liste des logiciels retenus et classification

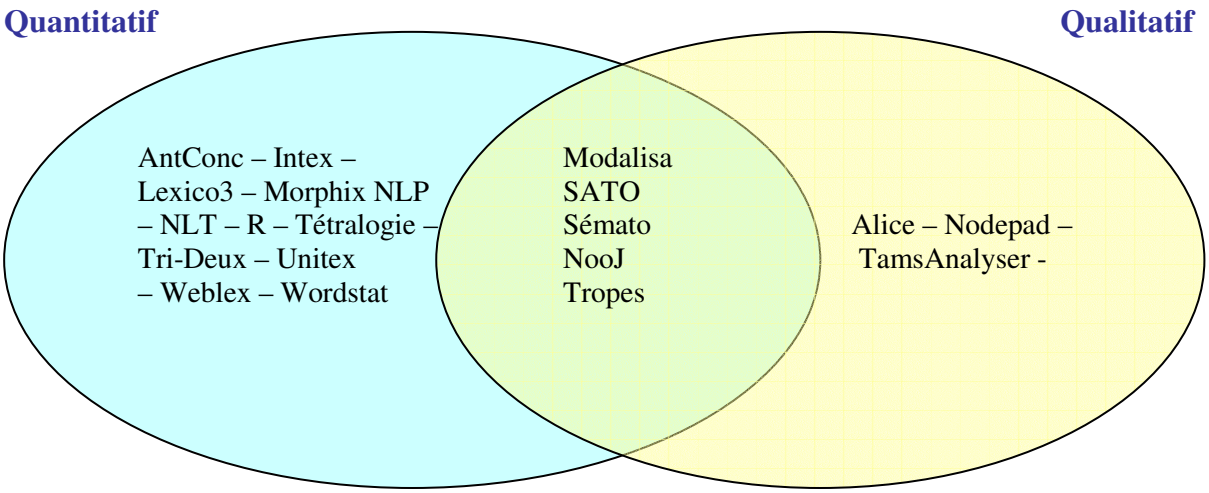
Ci-dessous figure la liste des logiciels retenus ainsi que la ou les catégories méthodologiques dont ils relèvent. Ces catégories ont été identifées sur base des descriptions trouvées, et confirmées par les tests pour les logiciels testés.

Logiciels	**Catégorie(s) méthodologique(s)**
Alice	Analyse socio-sémantique
AntConc	Lexicométrie, analyse par réseaux de mots associés
Atlas libre	Annotations
Intex	Lexicométrie
Lexico3	Lexicométrie, analyses statistiques, analyse par réseaux de mots associés
Modalisa (Interviews)	Lexicométrie, analyses statistiques, analyse socio-sémantique, analyse par réseaux de mots associés
Morphix NLP	Lexicométrie, analyses statistiques
NLT	Lexicométrie, analyses statistiques
Nodepad	Analyse socio-sémantique
Nooj	Analyse socio-sémantique
R	Lexicométrie, analyses statistiques
Sato	Lexicométrie, analyse socio-sémantique, analyse par réseaux de mots associés
Sémato	Analyse socio-sémantique, analyse par réseaux de mots associés
TamsAnalyser	Analyse socio-sémantique
Tetralogie	Lexicométrie, analyses statistiques, analyse par réseaux de mots associés
Transcriber	Annotations
Tri-Deux (Thèmes)	Lexicométrie, analyses statistiques
Tropes	Analyse par réseaux de mots associés, analyse socio-sémantique, analyse-cognitivo-discursive
Unitex	Lexicométrie
Weblex	Lexicométrie, analyses statistiques, analyse par réseaux de mots associés
Wordstat	Lexicométrie, analyses statistiques, analyse par réseaux de mots associés

Il convient également de noter que la plupart de ces logiciels intégrent d'autres fonctions telles la recherche de termes ou d'expressions avec affichage des concordances.

Lors du chapitre deux, les méthodes d'analyse informatisées avaient été classées selon différents critères. L'un de ces critères était le côté quantitatif ou qualitatif. Nous avions d'une part les méthodes quantitatives (lexicométrie, analyses statistiques et analyse par réseau de mots associés) et d'autre part les méthodes qualitatives (analyse propositionnelle du discours, propositionnelle prédicative et cognitivo-discursive), les analyses socio-sémantiques se plaçant à l'intersection des deux.

Le graphique ci-dessous montre la répartition des logiciels identifiés selon ce critère.



2. Présentation des certains logiciels

Dans ce point, nous allons présenter de manière approfondie quelques logiciels d'analyse de texte, l'idée étant d'en sélectionner un par méthode d'analyse afin de donner un aperçu large de différentes possibilités.

Les interviews des chercheurs qui seront présentées dans le chapitre 4 ont été réalisées avant le choix des logiciels à approfondir. Comme nous le verrons, les attentes variées des chercheurs nous ont poussé à choisir des logiciels contrastés afin de couvrir un maximum d'attentes.

Tropes étant le seul de sa catégorie, son choix s'est imposé de lui-même. Pour les autres, la présence d'un manuel détaillé permettant une prise en main rapide, ainsi que le correct déroulement des prétests ont présidé au choix. On donc été retenu : AntConc qui permet de faire de la lexicométrie et de l'analyse par réseau de mots associés ; Lexico représentant de la lexicométrie et des analyses statistiques textuelles ; Sémato pour l'analyse socio-sémantique et Unitex pour la lexicométrie basique (décompte des formes) et pour ses potentialités en matière de recherche textuelle.

En ce qui concerne le schéma d'analyse suivi pour présenter les logiciels, nous avons retenu les points suivants :

- Nous commencerons par quelques brèves généralités : concepteur, système d'exploitation, références pour la présentation ;
- Nous nous pencherons ensuite sur les fonctionnalités offertes par le logiciel : quel format accepte-il en entrée, quels sont les outils qu'il propose, quelles sorties (impressions, affichages, rapports, etc.) offre-t-il.
- Enfin, nous ferons une synthèse des éléments recueillis.

Chaque outil a été testé sur base d'un article de journal qui figure en annexe (Annexe 4).

2.1. ANTCONC[145]

a) Généralités

Le logiciel AntConc a été développé par Laurence Anthony de l'Université de Waseda au Japon. Il est décrit par son concepteur comme un concordanceur, disponible pour Linux, Windows et Mac. L'interface est en anglais.

La rédaction de cette présentation se base sur l'utilisation de l'outil AntConc et sur son manuel d'utilisation « Read me File for AntConc 3.1.303 (Windows and Linux) » du 31 juillet 2006 fourni avec le logiciel.

b) Fonctionnalités

b.1) Formats d'entrée

AntConc accepte uniquement des textes au format .txt.

b.2) Analyse du corpus et fonctions

Afin de pouvoir exécuter les fonctionnalités sur un ou plusieurs textes, il convient de « charger » ceux-ci dans le logiciel. Les noms de ceux-ci s'ajoutent alors dans la partie gauche (en gris, voir ci-dessous) de l'écran.

AntConc ne lance pas lui-même l'analyse du texte. Il appartient à l'utilisateur de choisir les fonctionnalités qu'il souhaite exploiter.

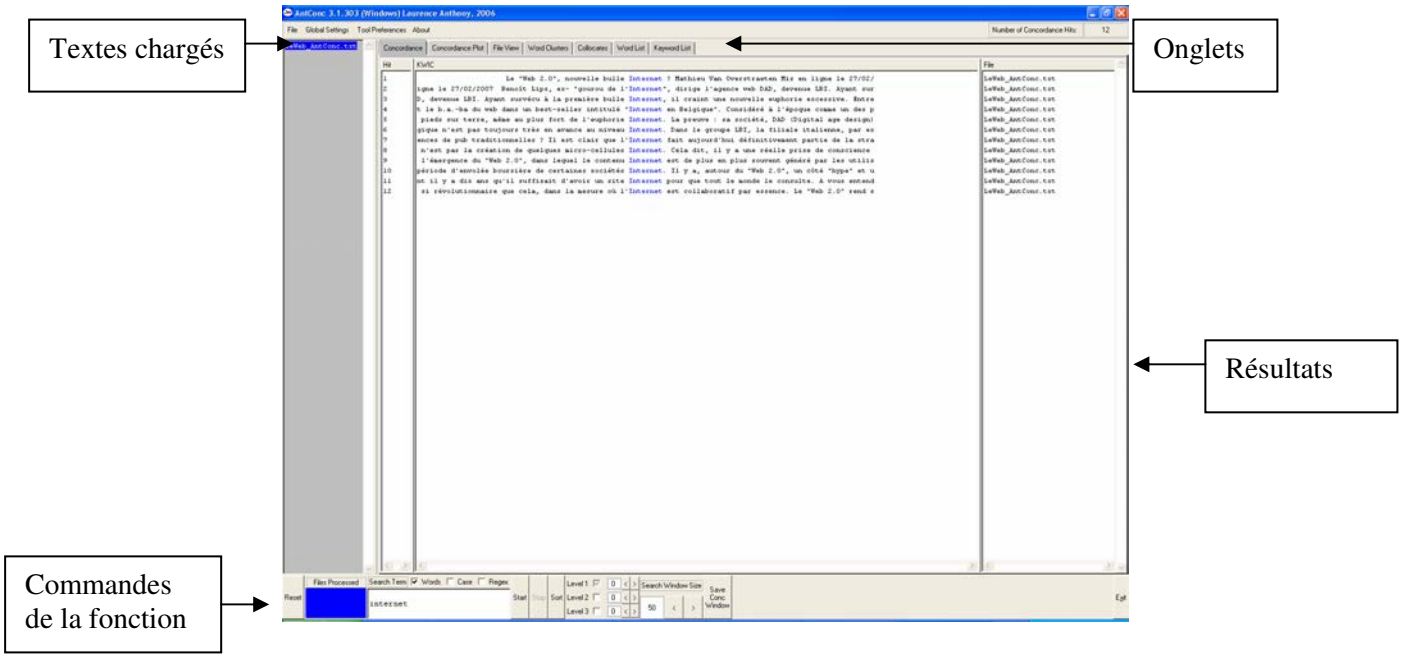
Les fonctionnalités vont être présentées selon l'ordre des onglets.

b.2.1) Concordance

L'outil « Concordance » permet de visualiser les contextes d'un terme précis, c'est-à-dire les phrases dans lesquelles il apparaît, dans un ou plusieurs textes.

Voici l'écran de résultat pour le terme « Internet ».

[145] http://www.antlab.sci.waseda.ac.jp/



Il est possible de personnaliser la taille du contexte.

Les résultats peuvent être triés en fonction du mot recherché (0), des termes de gauche (1L, 2L, …) ou des termes de droite (1R, 2R, …). Ce tri est possible sur trois niveaux.

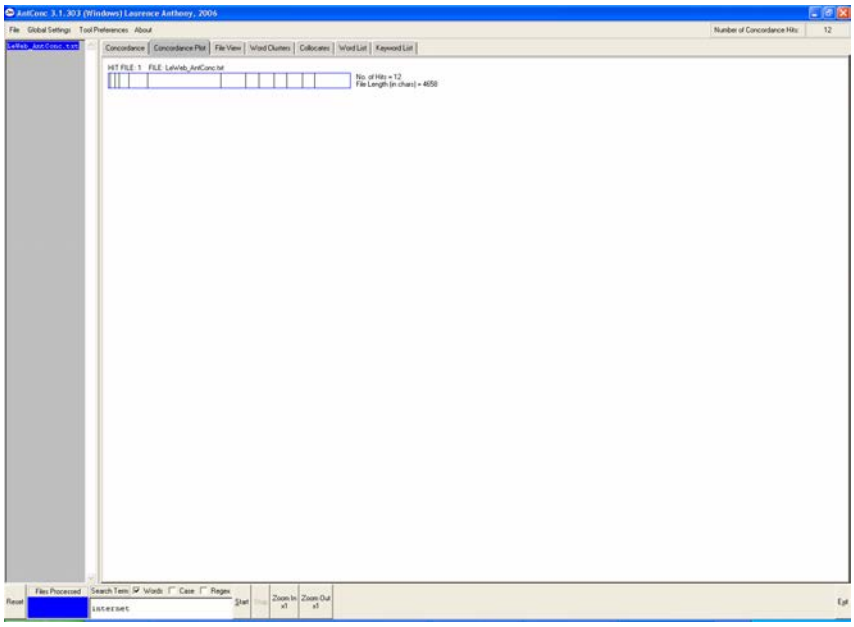
Il est possible de rechercher des termes ou des fragments de termes, de spécifier que la recherche est sensible à la casse, ou de rechercher des expressions régulières.

Les résultats affichent le numéro de la phrase, la phrase et le fichier qui la contient.

b.2.2) Concordance plot

L'outil « Concordance Plot » remplit la même fonction que l'outil « Concordance », mais via un affichage différent.

Voici le résultat pour le terme « Internet ».



Les résultats sont présentés sous la forme d'un code-barre dans lequel les traits indiquent la position des phrases contenant le terme recherché. La longueur du fichier ainsi que le nombre de résultats sont également indiqués.

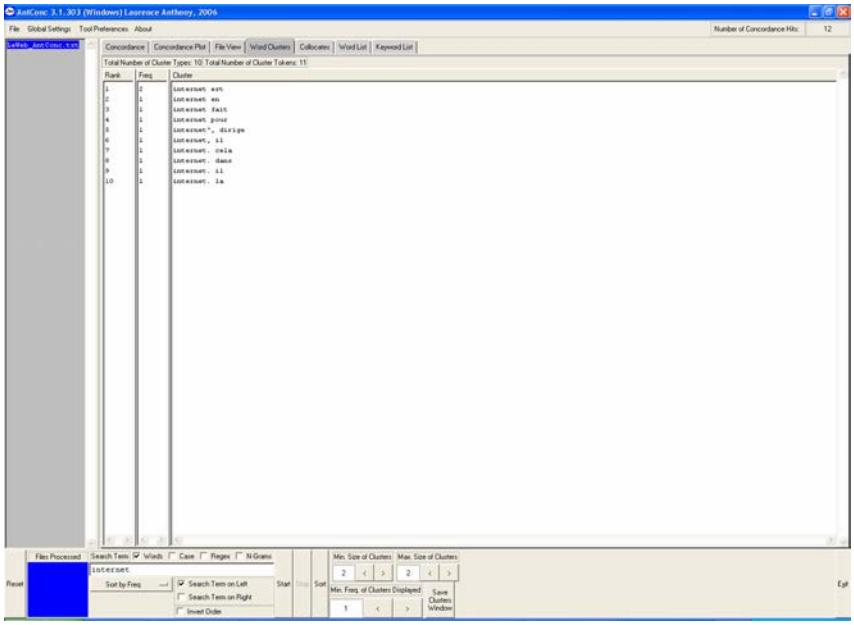
b.2.3) File view

L'outil « File view » permet de visualiser le texte en entier. Si des termes ont été sélectionnés pour une recherche, ils sont mis en évidence par l'utilisation de couleurs.

b.2.4) Word cluster

L'outil « Word cluster » permet de générer la liste des « clusters », c'est-à-dire des termes apparaissant juste après ou juste avant un terme spécifique.

Voici le résultat pour le terme « Internet ».



L'utilisateur doit spécifier s'il souhaite le cluster droit ou gauche, ou les deux.

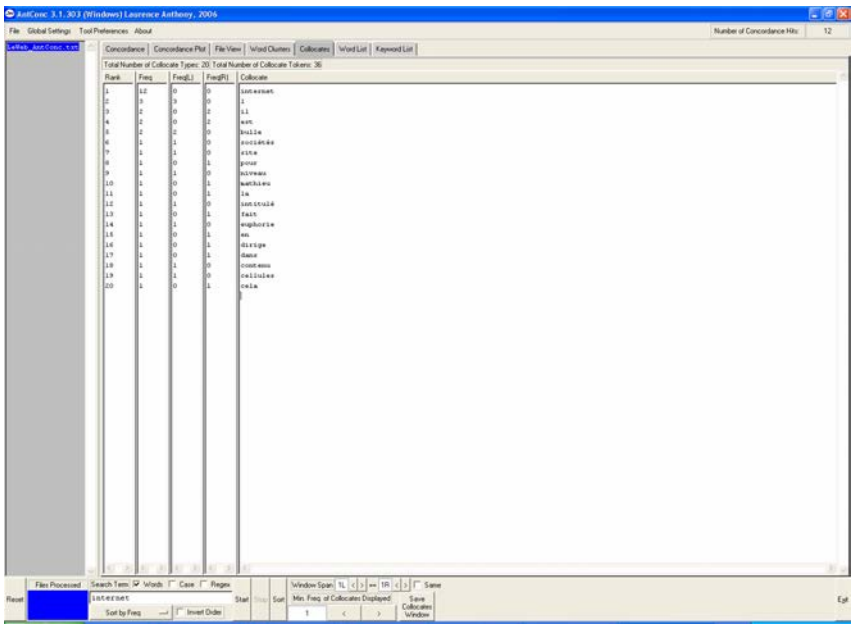
Il est possible de trier cette liste selon la fréquence, le début ou la fin du mot.
La taille minimale et maximale du cluster peut aussi être définie, ainsi que la fréquence minimale des clusters affichés.

Il est possible de rechercher des termes ou des fragments de termes, de spécifier que la recherche est sensible à la casse, ou de rechercher des expressions régulières ou des N-grammes.

Les N-grammes sont des N-grammes de mots dans le cadre de ce logiciel.
Par exemple, un N-gramme de taille 2 pour la phrase suivante « Le ciel est bleu » serait « Le ciel », « ciel est » ou « est bleu ».

b.2.5) Collocates

L'outil « Collocates » permet de visualiser les termes apparaissant dans le contexte d'un terme de base. Il s'agit en fait de rechercher les cooccurrences.



L'utilisateur peut spécifier la taille des contextes gauche et droit dans lesquels il faut rechercher les cooccurrences, et la fréquence minimale d'affichage des résultats.

A nouveau, cette fonction est applicable aux termes, aux segments de termes (avec sensibilité ou non à la casse) et aux expressions régulières.

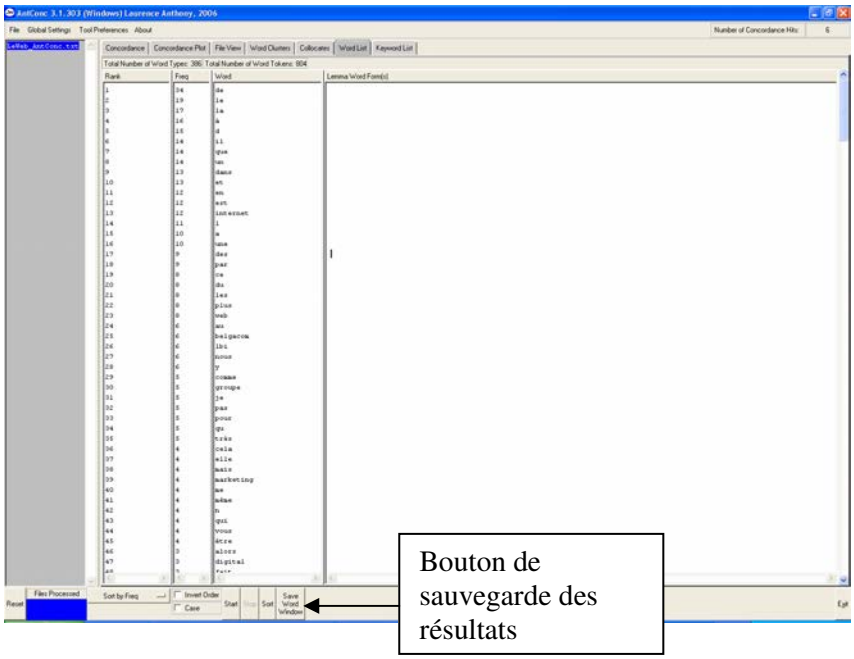
Il est possible de trier les résultats par fréquence, par fréquence du terme à droite ou à gauche, par le début ou la fin du terme, ou selon une mesure statistique de proximité[146] entre le terme recherché et ses cooccurrents.

[146] A savoir : (MI) Mutual Information: Using equations described in M. Stubbs, Collocations and Semantic Profiles, Functions of Language 2, 1 (1995) et (T-Score) T-Score: Using equations described in M. Stubbs, Collocations and Semantic Profiles, Functions of Language 2, 1 (1995)

b.2.6) Word List

L'outil « Word list » génère la liste des termes apparaissant dans le texte analysé ainsi que leur fréquence. Il procède donc à l'indexation de celui-ci. Celle-ci se base sur le repérage de caractères précis (lettres, chiffres, ponctuation, etc.). Par défaut, seules les lettres sont considérées comme formant des unités.

Voici les résultats pour le texte « Le Web ».



Ici aussi, l'on peut appliquer la sensibilité à la casse.

Les résultats peuvent être triés selon la fréquence, le début ou la fin d'un mot.

Il est intéressant d'utiliser cet outil en premier dans la mesure où il permet de sélectionner directement les termes sur lesquels on veut appliquer l'outil « Concordances ».

b.2.7) Keyword List

L'outil « Keyword List » permet de comparer le vocabulaire du texte étudié avec le vocabulaire d'un corpus de référence afin de générer une liste de mots-clés, exceptionnellement fréquents (ou non) dans le texte analysé.

b.3) Affichage et autres sorties

b.3.1) Affichage

L'affichage des résultats a déjà été illustré dans le cadre de la présentation des fonctionnalités.

Pour rappel, il se fait la plupart du temps sous forme de listes (termes, phrases, clusters, cooccurrences). Pour les concordances, un affichage sous forme de code-barre est également possible.

Il est généralement possible de trier les résultats selon différents critères exposés ci-dessus.

b.3.2) Sauvegarde des résultats

Pour chaque fonctionnalité, un bouton « Save XXX Windows » (dans la partie inférieure) permet de sauvegarder les résultats dans un fichier .txt.

c) Résumé des fonctionnalités

Au vu des fonctionnalités offertes, on peut classer le logiciel AntConc parmi les logiciels permettant de faire de la lexicométrie dans sa forme la plus simple (décompte des unités textuelles). Il peut également être classé parmi les outils apportant une aide pour faire de l'analyse par réseau de mots associés, c'est-à-dire rechercher des cooccurrences.

AntcConc permet de lancer des analyses sur plusieurs textes en même temps.
Ces analyses doivent toujours être lancées par l'utilisateur. AntConc n'effectue rien par lui-même.

En ce qui concerne les outils proposés,

- Il est possible de rechercher des « concordances » c'est-à-dire les contextes (phrases dans lesquelles ils apparaissent) d'un terme, d'un fragment de terme ou d'une expression régulière. Ces concordances peuvent être affichées sous forme textuelle ou sous forme d'un code-barre reprenant la position des phrases contenant l'objet de la recherche.
- Il est possible d'afficher le texte en entier avec mise en évidence de termes recherchés.
- Le logiciel permet de rechercher les « clusters » d'un terme, d'un fragment de terme, d'une expression régulière ou d'un N-gramme, c'est-à-dire les termes apparaissant juste avant ou juste après.
- L'outil « Collocates » permet de rechercher les cooccurrences de termes.
- L'outil « Word list » procède à une indexation des termes avec indication de leur fréquence.
- L'outil « Key Word List » permet la comparaison du vocabulaire du texte avec celui de textes de références.

Enfin, AntConc permet de sauvegarder des résultats dans un fichier .txt.

2.2. LEXICO3[147]

a) Généralités

Le logiciel Lexico3 a été développé par l'équipe universitaire SYLED-CLA2T de l'Université de la Sorbonne Nouvelle de Paris. Il est présenté comme un outil de lexicométrie et de statistiques textuelles.

La présentation du logiciel Lexico3 se base sur l'utilisation de celui-ci et sur son manuel « Lexico3 – Outil de statistiques textuelles – Manuel d'utilisation » (http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/).

[147] http://www.cavi.univ-paris3.fr/Ilpga/ilpga/tal/lexicoWWW/

b) Fonctionnalités

b.1) Formats d'entrée

Le logiciel Lexico3 accepte de fichiers textuels au format .txt uniquement.

Lexico3 permet d'introduire des balises ou clés dans le texte afin de séparer celui-ci en différentes parties.

Les balises doivent être introduites préalablement dans le texte sous le format suivant : <typebalise=nom>.

Par exemple, si l'on veut analyser les réponses de plusieurs personnes à une même question, on peut coder chaque réponse comme suit :

<question1=x> texte de la réponse
<question1=y> texte de la réponse
etc.

Lors de l'indexation, Lexico3 repère ces balises. En cas d'erreur de codage, il le signale à l'utilisateur qui doit corriger et relancer l'indexation.

b.2) Analyse du corpus et fonctions

Pour lancer l'analyse d'un texte, il faut cliquer sur cette icône  afin d'ouvrir le fichier à analyser. Lexico3 procède immédiatement à l'indexation du texte avec indication des fréquences.

L'indexation effectuée par Lexico3 se base sur une découpe en « forme graphique ».[148] Lexico3 n'utilise donc pas de dictionnaire.

Pour pouvoir reconnaître ces formes graphiques et segmenter le texte en mots, Lexico3 se base sur des caractères délimiteurs.

Lors du lancement d'une analyse, Lexico3 propose ces caractères via une boîte de dialogue : -—_:;/.,?¿!¡*$"+=(){}. Le « blanc » leur est automatiquement ajouté. Il est possible à l'utilisateur de modifier ceux-ci.

Certains caractères peuvent être choisis comme « délimiteur de section » afin d'affiner l'analyse du texte et de suivre sa structure originelle. Cela permet notamment de procéder à des comparaisons entre sections.

Tous les caractères qui ne sont pas repris dans les délimiteurs sont considérés comme non-délimiteurs et composent les formes graphiques.

[148] « une forme graphique est une suite de caractères non-délimiteurs, encadrée par deux caractères délimiteurs », « Lexico3 – Outil de statistiques textuelles – Manuel d'utilisation », p. 8.
D'autres découpes sont possibles : sur base des lemmes ou des n-grammes (racines).
La découpe en lemme se base sur des dictionnaires.

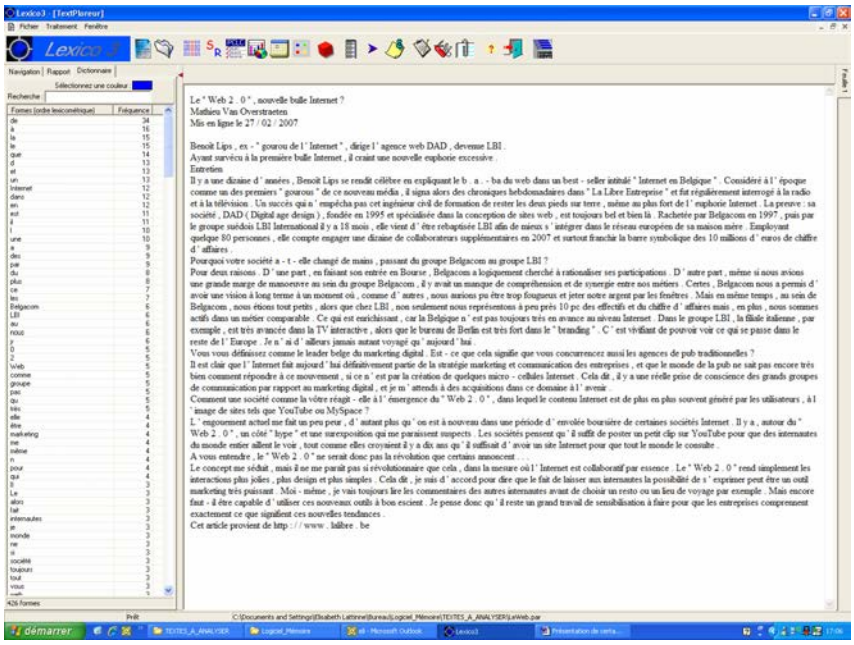
Pour être considérées comme égales et donc être comptées dans une même forme, deux formes graphiques doivent être absolument identiques.
Ainsi, la casse intervient. Les termes « informatique » et « Informatique » ne sont pas considérés comme identiques par Lexico3. De même, « c'est à dire » et « c'est-à-dire » seront considérés comme deux formes graphiques différentes.

Si l'utilisateur veut éviter ces inconvénients, il lui faut prétraiter son texte afin de détecter les sources de problèmes et d'y remédier[149].

D'autres unités textuelles que les formes graphiques sont toutefois identifiables par Lexico3 :

- Les segments[150] répétés : suites de formes graphiques identiques attestées plusieurs fois dans le texte, exemple : « langage de programmation orienté objet » ;
- Les cooccurrences : couples de formes présentes dans les mêmes contextes (phrase, sections, etc.) ;
- Les types généralisés ou Tgen(s): unités de dépouillement définies par l'utilisateur à l'aide d'outils lui permettant d'effectuer automatiquement des regroupements d'occurrences du texte, exemple : regroupement des singulier et pluriel, féminin et masculin (voir ci-après).

Le logiciel Lexico3 a été testé avec le texte « Le Web ». Voici les résultats affichés :



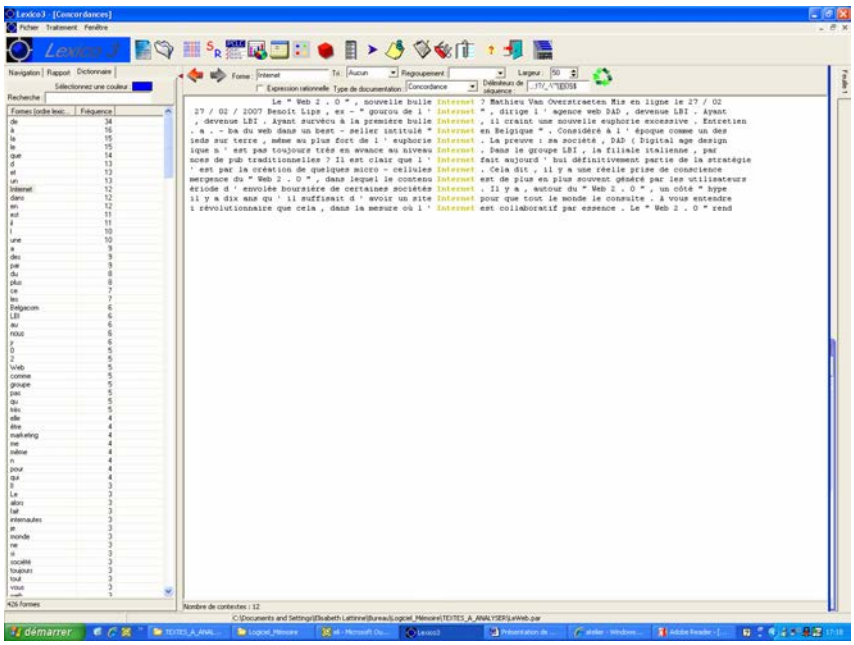
La partie gauche affiche la liste des formes graphiques présentes dans le corpus, munies de leur fréquence (indexation). Il est possible de les classer par fréquence ou alphabétiquement.
La partie droite affiche le texte analysé.

[149] Ainsi, afin de pouvoir considérer comme une même forme graphique, les termes en majuscule et minuscule, le manuel de Lexico3 propose de remplacer la forme majuscule par sa forme minuscule directement précédée du caractère *.
Ex. : Moi → *moi
Lors de l'analyse, selon que l'on veut les considérer ensemble ou non, il suffira d'introduire ou d'enlever le caractère * des délimiteurs.
[150] « toute suite d'occurrences consécutives dans le corpus et non séparées par un séparateur de séquence », « Lexico3 – Outil de statistiques textuelles – Manuel d'utilisation »

b.2.1) Concordance

L'outil « Concordance » permet de visualiser toutes les occurrences d'une forme ou d'un type généralisé en contexte, c'est-à-dire tous les morceaux de phrases dans lesquels ils apparaissent.

Par exemple, dans le cadre de l'analyse du texte « Le Web », l'outil a été appliqué au terme « Internet ».



Les contextes, phrases comprenant la forme « Internet », s'affichent à droite.

Il est possible de trier ces contextes selon différents critères :

- Avant : l'ordre alphabétique de l'occurrence qui précède la forme recherchée
- Après : l'ordre alphabétique de l'occurrence qui suit la forme recherchée
- Aucun : l'ordre d'apparition des occurrences de la forme recherchée dans le texte

Il est également possible de modifier la taille du contexte en modifiant la « largeur » (nombre de caractère avant et après la forme)

b.2.2) Segments répétés

Les segments répétés sont des suites de formes dont la fréquence est supérieure à 2 dans le corpus.

L'activation de cette fonction fait apparaître une boîte de dialogue qui permet de la paramétrer. Il est ainsi possible de déterminer quels caractères seront délimiteurs et le statut

des clés/balises rencontrées dans le texte. Il convient également de déterminer le seuil de sélection (fréquence minimale) des formes et des segments.

Le texte étant relativement court, aucun segment répété n'a été trouvé, même en modifiant le seuil.

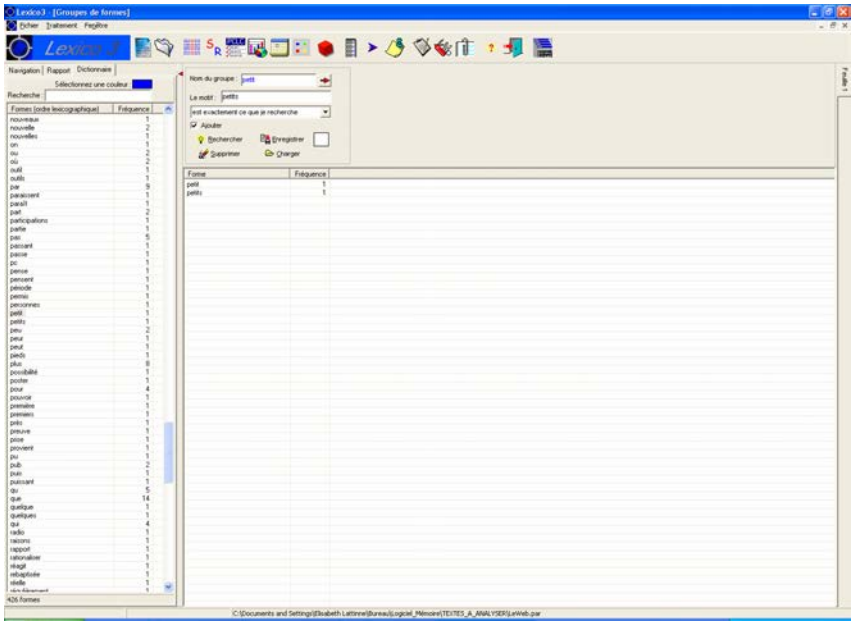
b.2.3) Groupes de formes

L'outil « Groupes de formes » permet de constituer des types rassemblant les occurrences de formes graphiques différentes liées par une propriété commune.

Cela peut servir, par exemple, à considérer ensemble le singulier et le pluriel d'un nom, ou les différentes formes conjuguées d'un verbe

Il suffit d'introduire le terme recherché, une partie de celui-ci (début, contenu, fin) ou une expression régulière. La partie droite inférieure affiche alors les formes graphiques répondant au critère de recherche.

Les termes non correspondant aux souhaits de l'utilisateur peuvent être supprimés de la liste.
Il est possible d'enregistrer les résultats dans un fichier dédié et de recharger celui-ci ultérieurement.



Dans cet exemple, on décide de regrouper « petit » et « petits ».

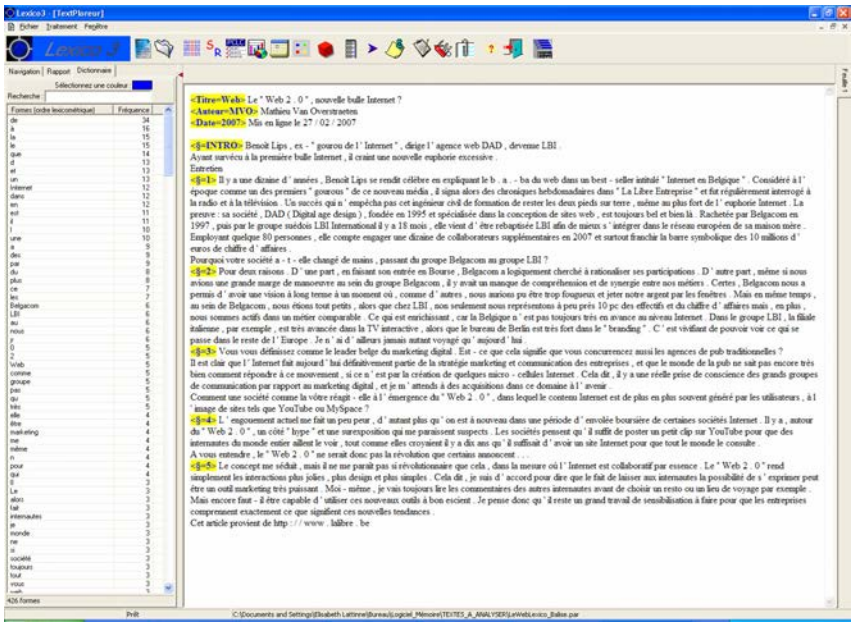
b.2.4) Outils d'analyse statistiques

Les outils d'analyse statistiques regroupent différentes méthodes qui seront décrites ci-après.

Leur utilisation présuppose un balisage préalable du texte et la sélection d'une clé.

La première analyse du texte « Le Web » avait été faite sans introduire de clé.
Le texte a été balisé par la suite afin de pouvoir présenter les analyses statistiques.
Il convient de souligner que les balises ont été décidées par le rédacteur de ce mémoire. Il apparaissait évident de mettre des clés spécifiques pour l'auteur, le titre et la date. Les paragraphes ont été décidés sur base du découpage du texte (retour à la ligne).

Voici l'écran de résultat affiché après le lancement et l'exécution de Lexico3.



On peut observer que les balises / clés introduites sont surlignées en jaunes.

1)) Statistiques générales et fréquences

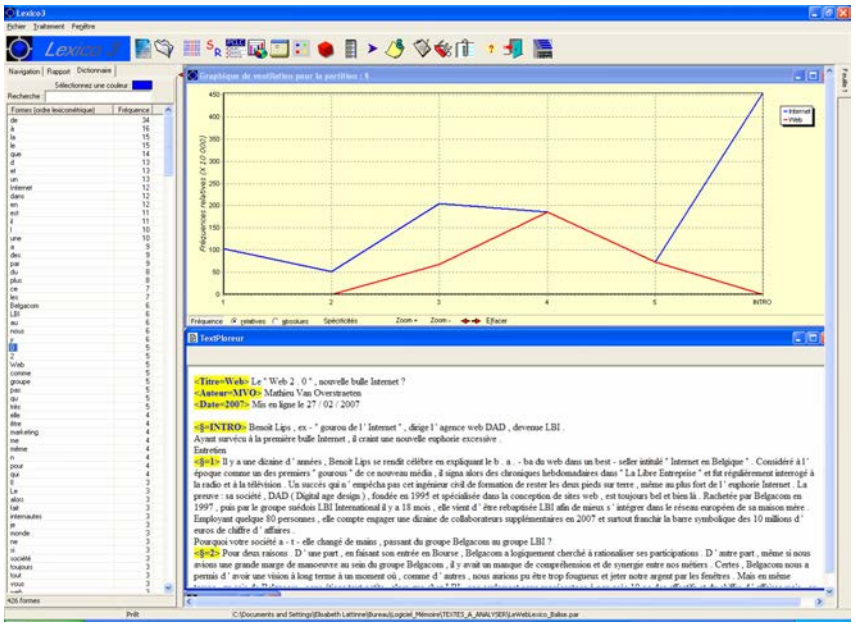
Le lancement de l'outil « Statistiques » fait apparaître une boîte de dialogue demandant à l'utilisateur la balise sur base de laquelle l'analyse statistique sera exécutée.

Pour l'exécution de l'outil « Statistiques », la clé « § » a été choisie pour des raisons évidentes.

L'écran de résultat généré permet à l'utilisateur de procéder à des comparaisons de formes, segments répétés ou types entre les différents paragraphes du texte.

Il suffit pour cela de glisser-coller la ou les formes graphiques que l'on veut étudier dans la partie jaune de la fenêtre.

Voici l'écran de résultat avec les termes « Internet » et « Web » (avec majuscule).

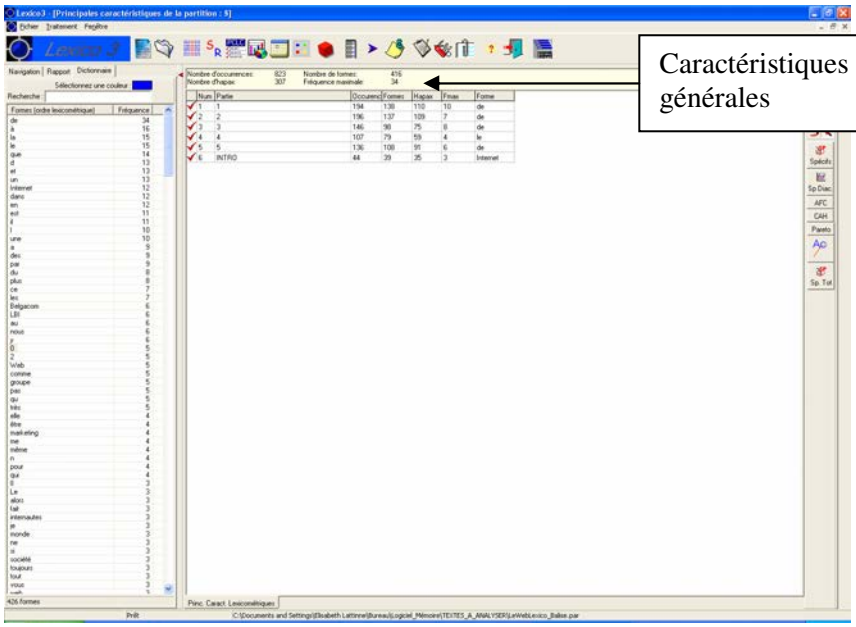


Cet écran montre la fréquence relative des formes « Internet » et « Web » dans les différents segments du texte (les § dans notre exemple).

La fréquence relative exprime le nombre d'occurrences rapportées à la longueur de la partie.
Il est également possible d'avoir les résultats en terme de fréquence absolue (nombre d'occurrences dans la partie) ou en terme de spécificités (résultat d'un calcul statistique, voir ci-dessous).

Cette fonction est disponible pour les formes, les segments répétés, les groupes de formes et ce qui aurait été mis dans le Garde-mots

L'activation de l'icône « Statistiques » a aussi pour effet de créer une autre fenêtre produisant des statistiques par parties et affichable en cliquant sur l'icône  .



La partie supérieure indique les caractéristiques générales du texte : nombre d'occurrences, nombre de formes, nombre d'hapax, fréquence maximale.

Dans la partie, inférieure, sont indiqués pour chaque partie : un numéro par partie (ici, 1 à 6), le nom de la partie, le nombre des occurrences des formes répertoriées, le nombre de formes graphiques identifiées, le nombre de formes qui n'apparaissent qu'une fois dans la partie, le nombre des occurrences de la forme la plus fréquente et la forme la plus fréquente.

2)) Spécificités

Sur la droite, le bouton « Specifs » permet d'obtenir un jugement sur la fréquence de chacune des unités textuelles dans chacune des parties du corpus.

Une présentation de cet outil statistique figure en annexe. En résumé, le calcul de la spécificité permet de déterminer si un terme est sur- ou sous-représenté dans les différentes parties du texte.

Pour appliquer cet outil, il est demandé à l'utilisateur d'indiquer le seuil de probabilité et la fréquence minimale des formes dont on veut tenir compte.

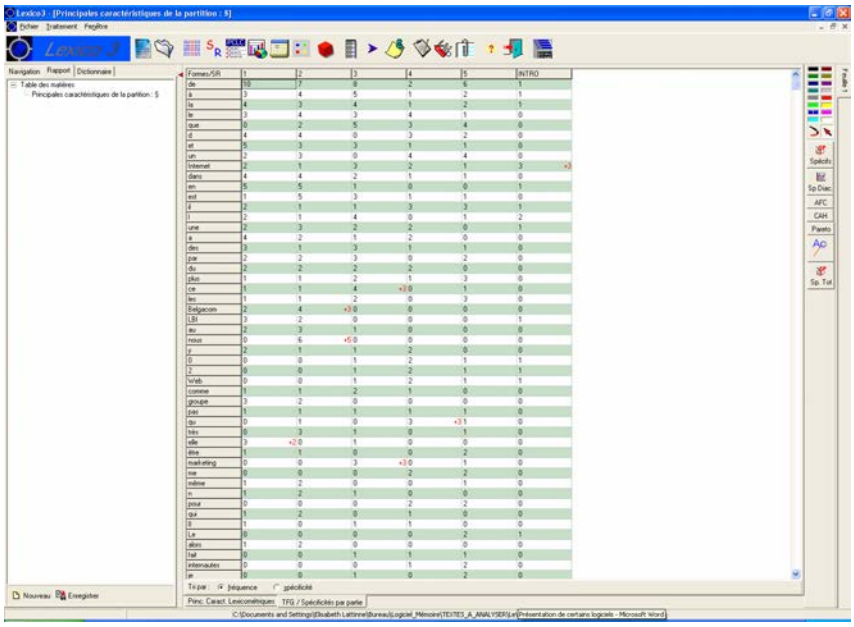
L'application de cet outil sur le texte « Le Web » n'a malheureusement pas donné de résultats, le texte devant être trop petit.

3)) Spécificités chronologiques

L'analyse des « spécificités chronologiques » met en évidence le vocabulaire particulier de périodes plus larges formées de parties consécutives pour les séries textuelles chronologiques (série de textes produits par une même source textuelle et régulièrement espacés dans le temps).[151]

En bref, cette fonctionnalité est intéressante si le corpus peut être découpé de manière temporelle. On peut prendre comme exemple un corpus composé de courriers. Une nouvelle balise serait introduite pour chaque nouvelle lettre. L'analyse des spécificités chronologique permettra de voir l'évolution du vocabulaire dans le temps.

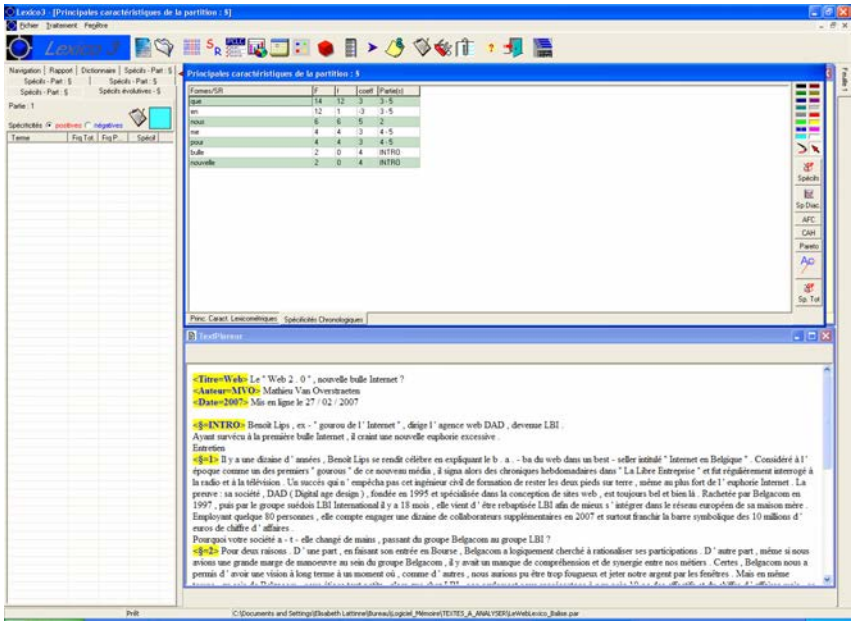
[151] « Lexico3 – Outil de statistiques textuelles – Manuel d'utilisation »



4)) Accroissements spécifiques

Il est aussi possible de calculer des « accroissements spécifiques » (bouton « Spec Diac »).

L'accroissement spécifique est une « spécificité calculée pour une partie d'un corpus par rapport à une partie antérieure ».



5)) Analyse factorielle des correspondances

Le bouton « AFC » permet d'effectuer une analyse factorielle des correspondances.

L'utilisateur est amené à paramétrer cette analyse :

- Le nombre des unités textuelles prises en compte dans l'analyse ;
- Le nombre des facteurs à extraire ;

- Le seuil de fréquence minimale.

Une erreur n'a pas permis d'obtenir des résultats.

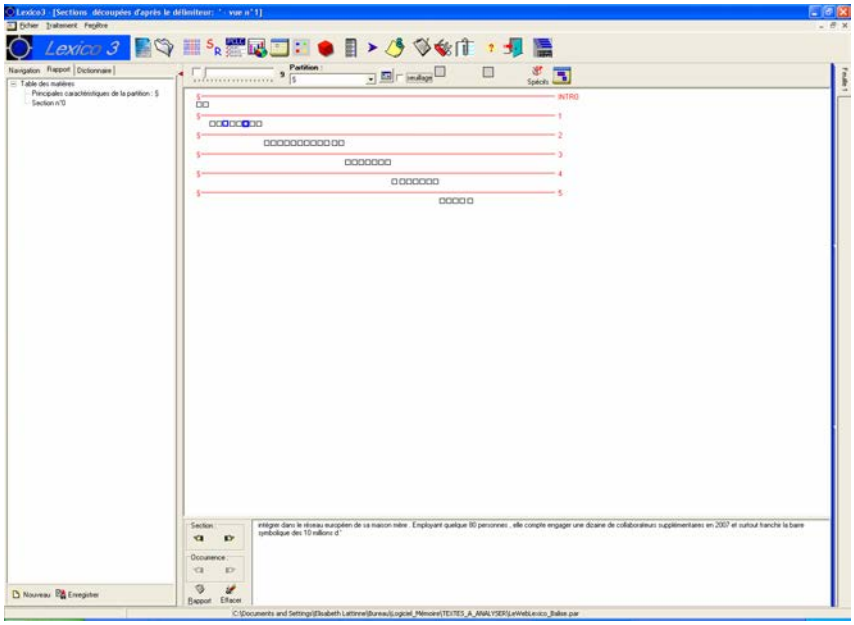
6)) Fonctions non décrites

Certains outils ne sont pas décrits dans le manuel, ce qui ne permet pas de les présenter. Il s'agit des fonctions « Pareto » et « Accroissement de vocabulaire ».

b.2.5) Outils de navigation

1)) Carte des sections

La « carte des sections » permet de visualiser le corpus découpé en sections lorsque des caractères particuliers (paragraphes, point, etc.) ont été choisis comme délimiteurs de section.



2)) Feuilles de travail

 et 

Les feuilles de travail permettent de d'éviter un fractionnement trop important de la fenêtre de travail principal. L'utilisateur peut en créer de nouvelles afin de répartir les résultats sur plusieurs feuilles.

b.3) Affichages et autres sorties

b.3.1) Fichiers de sortie

Au terme de l'analyse, plusieurs fichiers sont produits par Lexico3 : .par, .dic, .num.
Ceux-ci sont placés dans le même dossier que celui où figure le texte analysé.

Le fichier « .par » contient les principaux décomptes portant sur les formes, les occurrences, etc., ainsi que le rappel des caractères délimiteurs choisis lors de la segmentation.

Le fichier « .dic » contient le dictionnaire des formes classées par fréquence.

Ces deux fichiers sont consultables.

Le fichier « .num » contient le texte numérisé, c'est-à-dire sous une forme codée de façon compacte. Ce fichier à usage interne n'est pas consultable.

Il existe également un fichier atrace.txt contenant un rapport détaillé des opérations effectuées par le programme. En cas d'échec du traitement (par exemple, en cas d'erreur dans les balises), il peut être consulté afin d'en identifier les causes.

b.3.2) Rapport

Les résultats produits par les diverses fonctionnalités peuvent être ajoutés au « Rapport ».
Il s'agit d'un dossier manipulable à l'aide d'un navigateur Web et contenant un fichier index.htm.

On trouve le dossier « Rapport » dans le dossier « Lexico3 » créé par l'installation du logiciel.

b.4) Absence de dictionnaires

Contrairement à d'autres logiciels, Lexico3 n'est pas équipé de dictionnaires afin de réaliser ses analyses. L'identification des unités lexicales se fait sur base de caractères délimiteurs spécifiés par l'utilisateur.

c) Résumé

Lexico est donc un logiciel à classer parmi les outils permettant de faire de la lexicométrie et certaines analyses statistiques sur du texte.

Lexico permet de découper un texte via l'introduction de balises. Ce découpage est d'ailleurs obligatoire si l'on veut utiliser les outils statistiques.

Au niveau des fonctionnalités,

- Lexico procède à une indexation automatique du texte lors de l'ouverture de celui-ci avec indication des fréquences. Cette indexation se base sur le repérage de caractères délimiteurs.
- L'outil « concordance » permet d'afficher les contextes dans lesquels apparaisse l'objet d'une recherche (terme précis ou type généralisé).

Le logiciel permet la recherche des segments répétés, c'est-à-dire de suites de formes dont la fréquence est supérieure à deux.

- L'outil « Groupe de forme » permet de créer des types généralisés par le regroupement de termes proches (singulier-pluriel, formes conjuguées).
- Il offre différents outils statistiques : statistiques générales sur le texte, fréquence relative, fréquence absolue, spécificités, dans tout le texte ou par partie, analyse factorielle des correspondances, etc.

Lexico est équipé d'un garde-mot permettant de sauvegarder des objets de recherche.

Il permet également d'ajouter des résultats à un dossier « Rapport », contenant un fichier htm.

2.3.UNITEX[152]

a) Généralités

Unitex est un logiciel développé par Sébastien Paumier à l'Institut Gaspard-Monge de l'Université de Marne-la-Vallée (France).

La présentation de cet outil se base sur l'utilisation de celui-ci et sur le manuel d'utilisation « UNITEX 1.2 - MANUEL D'UTILISATION » de Sébastien Paumier (mai 2006).

Unitex est un logiciel libre. Il peut être utilisé sous Windows, Linux[153] et MacOs.

Il peut traiter des textes dans différentes langues (français, anglais, grec, italien, espagnol, allemand, thaï, coréen, polonais, norvégien, portugais, etc.).

b) Fonctionnalités

b.1) Formats d'entrée

Unitex manipule des textes Unicode. Donc, si le fichier à analyser n'est pas dans ce format, il en propose la conversion ou la création d'une copie du fichier source au bon format.

En ce qui concerne les types de fichier ouvrables par Unitex, ils sont de deux formats : .txt et .snt.
Les fichiers portant l'extension .snt sont des fichiers textes prétraités par Unitex qui sont prêts à être manipulés par les différentes fonctions du système. Les fichiers portant l'extension .txt sont des fichiers textes bruts.

b.2) Analyse du corpus et fonctions

b.2.1) Ouverture d'un fichier pour analyse

Lors de l'ouverture d'un fichier .txt (le texte « Le Web » pour l'exemple), Unitex demande à l'utilisateur s'il souhaite prétraiter celui-ci. Le prétraitement consiste à effectuer les opérations suivantes : normalisation des séparateurs, découpage en unités lexicales, normalisation de formes non ambiguës, découpage en phrases et application des dictionnaires.

Les séparateurs sont l'espace, la tabulation et le retour à la ligne. Les normaliser revient à effectuer les opérations suivantes :

[152] http://www-igm.univ-mlv.fr/~unitex/
[153] Unitex est utilisable en ligne de commande. Le manuel reprend l'ensemble des programmes compris dans Unitex et les commandes associées. Ces programmes se trouvent dans le répertoire Unitex/App.

- toute suite de séparateurs contenant au moins un retour à la ligne est remplacée par un unique retour à la ligne ;
- toute autre suite de séparateurs est remplacée par un espace.

Le découpage en unités lexicales est dépendant de la langue.

La normalisation des formes non ambiguës s'applique à des formes telles « l'on » qui peut être remplacé par « on ». Elle se base sur l'utilisation de grammaires.

Le découpage en phrase s'effectue sur base de grammaires qui décrivent les différents contextes où peuvent apparaître les limites de phrases, afin de solutionner les ambiguïtés.

L'application de dictionnaires au corpus conduit à l'indexation de celui-ci. Elle produit une fenêtre de résultat appelée « Word List » ne contenant que les formes présentes dans le texte (voir ci-après).

Le choix de prétraiter le texte fait apparaître une boîte de dialogue contenant différentes options.[154]

En cas de refus du prétraitements, le texte est néanmoins normalisé et découpé en unités lexicales.

Différentes fenêtres sont créées.
Une première fenêtre comprend le texte soumis au prétraitement. L'opération de prétraitement produit également une fenêtre « Token List » contenant toutes les unités lexicales du texte prétraité, ainsi que leur fréquence. En fin, la fenêtre « Word List » est divisée en trois parties :

- dlf : contient les mots simples ;
- dlc : contient les mots composés ;
- err : contient les mots inconnus ;

triés par ordre alphabétique.

texte

token list

word list

[154] L'option « Apply grah in MERGE mode » sert à effectuer le découpage du texte en phrases.
L'option « Apply graph in REPLACE mode » est utilisée pour effectuer des remplacements dans le texte, le plus souvent des normalisations de formes non ambiguës.
L'option « Apply All default Dictionaries » permet d'appliquer au texte des dictionnaires au format DELA[154].
L'option « Construct Text Automaton » est utilisée pour construire l'automate du texte.
« Cancel but tokenize text » permet d'effectuer uniquement la normalisation des séparateurs et le découpage en unités lexicales.

b.2.2) Recherche d'expressions rationnelles

L'outil de recherche d'Unitex est une des principales fonctionnalités de celui-ci.

Il permet de faire des recherches des plus simples aux plus complexes dans le corpus en utilisant des expressions rationnelles.

Dans le cadre d'Unitex, une expression rationnelle peut-être :

- une unité lexicale (les mots, par exemple « livre ») ;
- un motif spécial[155] ou un masque lexical (informations apparaissant dans les dictionnaires, par exemple <manger.V>) ;
- la concaténation de deux expressions rationnelles (les deux termes doivent être présents, par exemple « je mange » qui équivaut à « je » + « mange ») ;
- l'union de deux expressions rationnelles (l'un des deux termes doit être présent, par exemple « Pierre+Paul » qui équivaut à « Pierre » ou « Paul ») ;
- l'étoile de Kleene d'une expression rationnelle (0, 1 ou plusieurs répétitions du termes, par exemple « très* ») ;
- des expressions régulières au format POSIX.

Différentes options sont offertes, notamment la possibilité de donner priorité aux séquences les plus longues ou les plus courtes, de limiter la recherche à un certain nombre d'occurrences, de déterminer le mode de tri des résultats, de sélectionner la longueur en caractères des contextes gauche et droit des occurrences qui seront affichées dans la concordance, de construire un fichier texte avec toutes les phrases contenant les occurrences.

La sélection d'une occurrence ouvre la fenêtre du texte et y sélectionne la séquence reconnue. Si l'automate du texte est construit, l'automate de la phrase contenant l'occurrence cliquée est chargé également.

b.2.3) Automate du texte - Construct FST-text

Une autre fonctionnalité importante d'Unitex est la fonction « Construct FST-text » qui permet de construire l'automate du texte. Ce dernier représente toutes les phrases du texte analysé, les chemins représentés exprimant toutes les interprétations possibles.
Pour chaque unité lexicale, toutes les interprétations possibles sont exprimées.

Cet outil est utile pour la levée des ambiguïtés.

Il est recommandé d'avoir découpé le texte en phrases et de lui avoir appliqué les dictionnaires avant d'effectuer cette opération. Si le texte n'est pas découpé en phrases, le

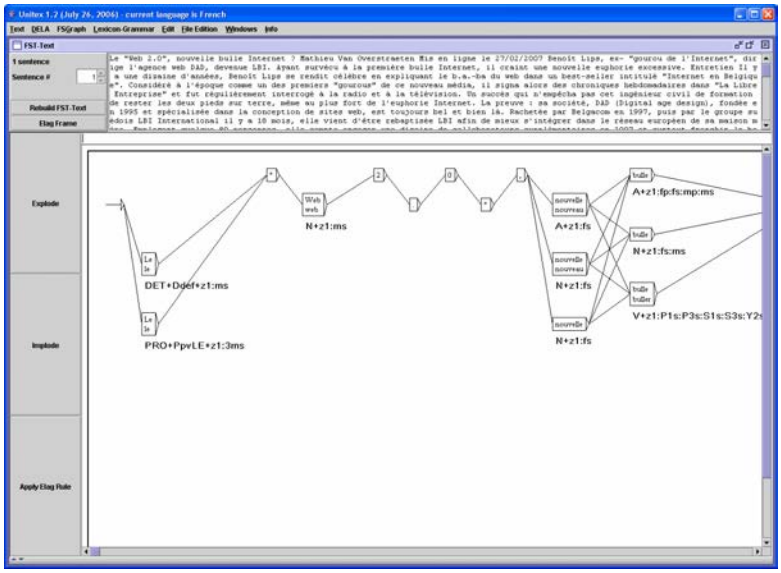
[155] Les motifs spéciaux sont les suivants :

- <E> : mot vide, ou epsilon. Reconnaît la séquence vide ;
- <TOKEN> : reconnaît n'importe quelle unité lexicale ;
- <MOT> : reconnaît n'importe unité lexicale formée de lettres ;
- <MIN> : reconnaît n'importe unité lexicale formée de lettres minuscules ;
- <MAJ> : reconnaît n'importe unité lexicale formée de lettres majuscules ;
- <PRE> : reconnaît n'importe unité lexicale formée de lettres et commençant par une majuscule ;
- <DIC> : reconnaît n'importe quel mot figurant dans les dictionnaires du texte ;
- <SDIC> : reconnaît n'importe quel mot simple figurant dans les dictionnaires du texte ;
- <CDIC> : reconnaît n'importe quel mot composé figurant dans les dictionnaires du texte ;
- <NB> : reconnaît n'importe quelle suite de chiffres contigus (1234 est reconnu mais pas 1 234) ;
- # : interdit la présence de l'espace.

programme découpe arbitrairement le texte en séquences de 2000 unités lexicales au lieu de construire un automate par phrase.

Les automates de phrase sont construits à partir des dictionnaires du texte. Le degré d'ambiguïté obtenu est donc directement lié à la finesse de description de ceux-ci.

Voici une partie de l'automate du texte « Le Web ».



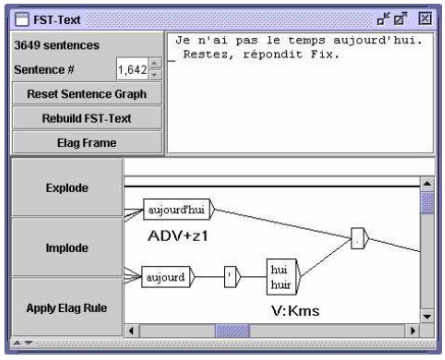
Chaque élément du texte est analysé. Certains termes qui peuvent être ambigus reçoivent plusieurs analyses. Dans notre exemple, « le » peut être un déterminant (DET) ou un pronom (PRO). De même, « nouvelle » peut être un adjectif (A) ou un nom (N).

Dans chaque boîte, la première ligne contient la forme fléchie (celle qui apparaît dans le texte) et la seconde la forme canonique (le lemme) s'il y a lieu d'être.

Dans le cadre de l'exemple, Unitex a considéré l'ensemble du texte comme une seule phrase. En cas de découpage en plusieurs phrase, il est possible de parcourir celles-ci individuellement.

Ce graphe peut être modifié par l'utilisateur (ajout ou suppression de boîtes, de liens, etc.).

Il est possible que dans certains cas, des mots inconnus parasitent l'automate. Par exemple, le terme « aujourd'hui » peut être analysé de deux manières par l'automate (exemple repris du manuel).



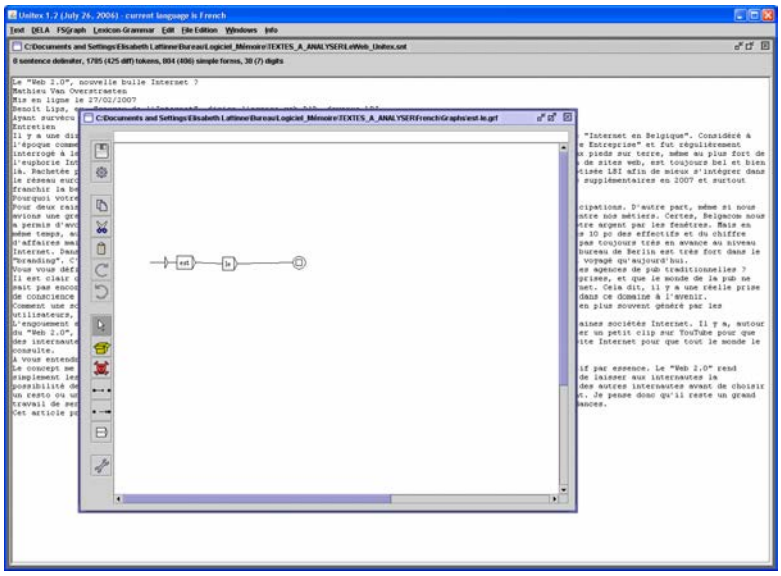
Il est possible de supprimer ces chemins parasites. Ce nettoyage s'effectue selon le principe suivant : si plusieurs chemins sont en concurrence dans l'automate, le programme garde ceux qui contiennent le moins de mots inconnus.

b.2.4) Grammaires locales

Unitex permet de créer des grammaires locales en utilisant des graphes.

L'objectif de ces grammaires est de rechercher des expressions dan le texte.
Le terme expression est utilisé dans un sens bien précis. Il désigne des phrases ou portions de phrases présentant des caractéristiques bien précises.

Par exemple, on peut vouloir rechercher touts les adjectifs suivis d'un nom commun, ou toutes les phrases de la forme : pronom personnel, verbe, adjectif.



Le symbole en forme de flèche est l'état initial du graphe.
Le symbole rond contenant un carré est l'état final du graphe. La grammaire ne reconnaîtra que les expressions décrites par des chemins reliant l'état initial à l'état final.
Les boîtes représentent les éléments de l'expression recherchée.

Les graphes ainsi créés peuvent être appliqués au corpus. Le résultat produit une concordance, c'est-à-dire un affichage de toutes les portions de phrases correspondant au graphe

b.3) Affichage et autres sorties

b.3.1) Répertoire personnel de travail et fichiers produits

Lors de la première utilisation sous Windows, Unitex demande à l'utilisateur de spécifier un répertoire de travail. Ce répertoire contiendra tous les fichiers produits par l'utilisation d'Unitex.

Lors des opérations effectuées par l'utilisateur ou automatiquement par Unitex, différents fichiers sont générés. Certains peuvent être consultés et édités.

b.3.2) Graphes

Il est possible de copier les graphes créés par l'utilisateur, dans d'autres applications en les enregistrant en tant qu'image au format PNG.

Il est également possible de les imprimer (graphes créés par utilisateur ou de l'automate) directement.

Les graphes de l'automate, non ambigus, peuvent être convertis en un fichier texte correspondant à l'unique chemin représenté par cet automate.

b.4) Présence de dictionnaires

L'exécution d'Unitex repose sur la présence de dictionnaires électroniques, de grammaires et de tables de lexique-grammaire.

Les dictionnaires électroniques décrivent les mots simples et composés d'une langue en leur associant un lemme ainsi qu'une série de codes grammaticaux, sémantiques et flexionnels.

Exemples :

mercantiles,mercantile.A+z1:mp:fp
gît,gésir.V+z1:P3s
grand mères,grand mère.N:fp
grand-mères,grand-mère.N:fp

Code	Signification	Exemples
A	adjectif	fabuleux
ADV	adverbe	réellement, à la longue
CONJC	conjonction de coordination	mais
CONJS	conjonction de subordination	puisque, à moins que
DET	déterminant	ses, trente-six
INTJ	interjection	adieu, mille millions de mille sabords
N	nom	prairie, vie sociale
PREP	préposition	sans, à la lumière de
PRO	pronom	tu, elle-même
V	verbe	continuer, copier-coller

codes grammaticaux

Code	Signification	Exemple
z1	langage courant	blague
z2	langage spécialisé	sépulcre
z3	langage très spécialisé	houer
Abst	abstrait	bon goût
Anl	animal	cheval de race
AnlColl	animal collectif	troupeau
Conc	concret	abbaye
ConcColl	concret collectif	décombres
Hum	humain	diplomate
HumColl	humain collectif	vieille garde
t	verbe transitif	foudroyer
i	verbe intransitif	fraterniser
en	particule pré-verbale (PPV) obligatoire	en imposer
se	verbe pronominal	se marier
ne	verbe à négation obligatoire	ne pas cesser de

codes sémantiques

Code	Signification
m	masculin
f	féminin
n	neutre
s	singulier
p	pluriel
1, 2, 3	$1^{ère}$, $2^{ème}$, $3^{ème}$ personne
P	présent de l'indicatif
I	imparfait de l'indicatif
S	présent du subjonctif
T	imparfait du subjonctif
Y	présent de l'impératif
C	présent du conditionnel
J	passé simple
W	infinitif
G	participe présent
K	participe passé
F	futur

codes flexionnels

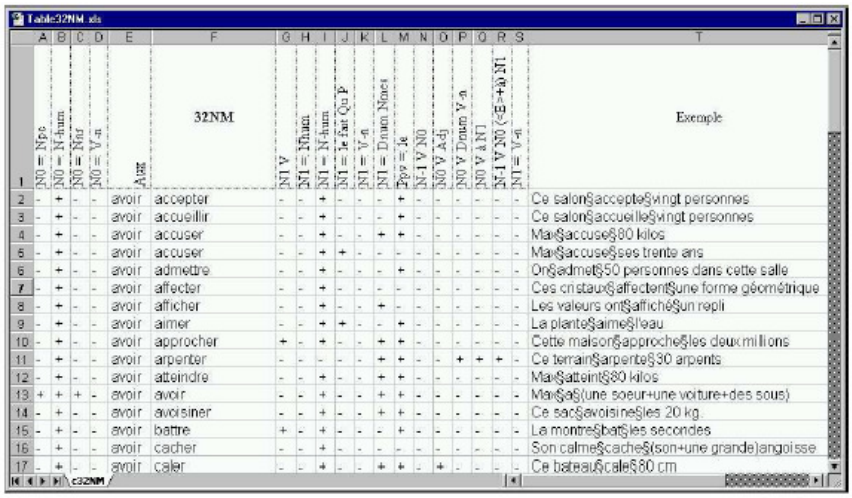
Ces dictionnaires sont consultables et éditables.

Les grammaires sont des représentations de phénomènes linguistiques par réseaux de transitions récursifs, formalisme proche de celui des automates à états finis.
Ces grammaires sont représentées au moyen de graphes éditables.

Les tables de lexique-grammaire, réalisées à l'aide de tableurs, sont des matrices décrivant les propriétés de certains mots.
Ces tables permettent de donner la grammaire de chaque élément de lexique, d'où le nom de lexique-grammaire.

Dans ces matrices, les lignes correspondent aux verbes et les colonnes aux propriétés syntaxiques.

Les propriétés considérées sont des propriétés formelles telles que le nombre et la nature des compléments admis par le verbe et les différentes transformations que ce verbe peut subir (passivation, nominalisation, extraposition, etc.). Les matrices sont binaires : un signe + apparaît à l'intersection d'une ligne et d'une colonne d'une propriété si le verbe vérifie la propriété, un signe - sinon.



	32NM	Aux	N0 = Npc	N0 = N-hum	N0 = Nar	N0 = V-n	N1 V	N1 = Nhum	N1 = N-hum	N1 = le fait Qu P	N1 = V-n	N1 = Dnum Nnc	Ppv = de	N-1 V N0	N0 V A-dj	N0 V Dnum V-n	N0 V à N1	N-1 V N0 (=E=+à) N1	N1 = V-n	Exemple
2	accepter	avoir	-	+	-	-	-	-	+	-	-	-	+	-	-	-	-	-	-	Ce salon§accepte§vingt personnes
3	accueillir	avoir	-	+	-	-	-	-	+	-	-	-	+	-	-	-	-	-	-	Ce salon§accueille§vingt personnes
4	accuser	avoir	-	+	-	-	-	-	+	-	-	+	+	-	-	-	-	-	-	Mar§accuse§80 kilos
5	accuser	avoir	-	+	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	Mar§accuse§ses trente ans
6	admettre	avoir	-	+	-	-	-	-	+	-	-	-	+	-	-	-	-	-	-	On§admet§50 personnes dans cette salle
7	affecter	avoir	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	Ces cristaux§affectent§une forme géométrique
8	afficher	avoir	-	+	-	-	-	-	+	-	-	+	-	-	-	-	-	-	-	Les valeurs ont§affiché§un repli
9	aimer	avoir	-	+	-	-	-	-	+	+	-	-	+	-	-	-	-	-	-	La plante§aime§l'eau
10	approcher	avoir	+	-	+	-	-	-	+	-	-	+	+	-	-	-	-	-	-	Cette maison§approche§les deux millions
11	arpenter	avoir	-	-	-	-	-	-	+	-	-	+	+	-	-	+	+	+	-	Ce terrain§arpente§30 arpents
12	atteindre	avoir	-	-	+	-	-	-	+	-	-	+	+	-	-	-	-	-	-	Mar§atteint§80 kilos
13	avoir	avoir	+	+	+	-	-	-	+	-	-	+	+	-	-	-	-	-	-	Mar§a§(une soeur+une voiture+des sous)
14	avoisiner	avoir	-	+	-	-	-	-	+	-	-	+	+	-	-	-	-	-	-	Ce sac§avoisine§les 20 kg
15	battre	avoir	-	+	-	-	+	-	+	-	-	-	+	-	-	-	-	-	-	La montre§bat§les secondes
16	cacher	avoir	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	Son calme§cache§(son+une grande)angoisse
17	caler	avoir	-	+	-	-	-	-	+	-	-	+	+	-	+	-	-	-	-	Ce bateau§cale§80 cm

c) Résumé

Unitex est un logiciel qui peut être classé dans les logiciels permettant d'effectuer de la lexicométrie basique, c'est-à-dire de procéder à l'indexation des unités lexicales avec indication des fréquences.

A côté de cette fonctionnalité, Unitex offre certains outils particuliers par rapport aux autres logiciels :

- Il offre différents outils de recherche permettant d'effectuer des recherches simples comme très complexes dans le texte : termes, expressions régulières, motifs, expressions.
- Il permet de construire le graphe du texte, c'est-à-dire de produire un graphique dans lequel chaque unité lexicale est analysée syntaxiquement et où les différentes interprétations sont indiquées s'il y a lieu.

2.4. SEMATO

a) Généralités

Sémato est un logiciel d'analyse de texte en ligne disponible à l'adresse suivante : http://fable.ato.uqam.ca/guidexpert-ato/gea.asp.

Il a été développé par l'Université du Québec à Montréal.

Il en existe deux versions : mode simple et mode avancé. Il peut analyser des textes français ou anglais.

La prise en main du logiciel s'est basée sur l'utilisation de la version « mode simple », sur le tutoriel et sur le manuel d'utilisation disponible en ligne (http://fable.ato.uqam.ca/guidexpert-ato/manuel-normal.asp).

b) Fonctionnalités

b.1) Formats d'entrée

L'utilisation de Sémato commence par la création d'un projet et l'activation de celui-ci via le lien « Projet » de la page d'accueil.

La création d'un projet se fait via « Ouverture d'un projet ». Il est demandé à l'utilisateur un nom pour le projet, un mot de passe et une adresse mail.

Une fois le projet accepté, celui-ci est activé. L'activation crée un cookie sur l'ordinateur de l'utilisateur, ce qui lui permet de revenir ultérieurement travailler sur son projet.
Lors de sessions ultérieures, l'activation du projet permet de récupérer celui-ci en cas de destruction du cookie. Elle permet également de naviguer entre les différents projets.

La caractéristique de Sémato est de permettre l'analyse de corpus qui sont transférés sur son serveur. L'utilisateur se voit donc réserver un espace personnel de travail.

Le transfert d'un corpus peut se faire selon différents modes : « Entrée express » ou « Génération de questionnaires Web ».

Le mode « Entrée express » permet de préparer et de transférer les données textuelles et extra-textuelles (diverses catégories) d'un projet.

Les « données extra-textuelles » correspondent à des catégories que l'utilisateur veut associer à ses données textuelles. Par exemple, en cas d'interviews, l'on veut pouvoir indiquer l'auteur, son âge, son sexe, etc. Le transfert de ces données est facultatif.

Le « Transfert de données textuelles » ouvre un formulaire permettant à l'utilisateur d'introduire son corpus, ainsi que certaines caractéristiques : nom du texte, repère de questions, repère de réponses (facultatif, en cas de questionnaire[156]).

[156] Par exemple, l'utilisateur pourrait indiquer Q au début des questions et R au début des réponses. L'indication de ces caractères dans les repères permettra à Sémato d'en tenir compte. Sémato va alors ajouter aux unités textuelles une catégorie de projet appelée TypeQR dont les valeurs possibles sont Question ou Réponse.

L'utilisateur doit répéter l'opération pour chaque texte de son corpus s'il y en a plusieurs. En effet, il n'est pas possible d'importer plusieurs textes à la fois.

Une fois le transfert lancé, Sémato procède à l'analyse du texte et découpe celui-ci en unités textuelles, afin de produire une base de connaissance pour les analyses ultérieures. Cette opération correspond à la fonction d'« indexation » qui sera détaillée ci-après.
Une unité textuelle au sein de Sémato est définie par un retour chariot. Elle correspond à ce qu'on appelle généralement un paragraphe dans le langage courant.
Ces unités sont appelées « textes » au sein de Sémato, ce qui peut être déroutant car ce terme est généralement employé pour un ensemble de paragraphes.
Il appartient donc à l'utilisateur de « découper » préalablement son corpus comme bon lui semble en introduisant les retours chariot adéquats.

Le transfert de données extra-textuelles suppose le transfert d'une table (excel) reprenant les méta-catégories souhaitées. Cette fonctionnalité n'a pas été testée.

Une autre manière d'introduire un corpus est le mode « Génération de questionnaires Web ».

Ce mode permet de générer un questionnaire WEB afin de recueillir des données textuelles via Internet. Il s'agit d'indiquer dans la fenêtre de saisie les directives pour la construction du questionnaire et de les soumettre. Sémato répond alors en donnant une adresse WEB correspondant au questionnaire.

Chaque fois qu'un répondant remplit le questionnaire, un fichier est créé dans le dossier INPUT du dossier projet.

Cette fonctionnalité n'a pas non plus été testée.

b.2) Fonctions

Sémato propose quatre fonctionnalités / outils principaux : Indexation, Requête, Thèmes et Pages d'arrimage.

La fonction « Indexation » a pour objectif de préparer le texte afin de permettre les analyses ultérieures.

Les fonctions « Requête » et « Thèmes » permettent ces analyses proprement dites.

L'outil « Pages d'arrimage » permet au chercheur d' « attacher » des thèmes ou mots clés aux différentes parties du corpus.

b.2.1) Indexation

La fonction d'« Indexation », préalable à toute opération ultérieure, procède à l'organisation de la base de connaissance du corpus (découpage en unité textuelle, voir ci-dessus).

Il est rappelé que l'utilisateur peut introduire des caractéristiques pour nommer son texte. Il lui est également demandé d'indiquer la langue du corpus.

Une fois la soumission opérée. Une page d'information est affichée. Elle donne divers renseignements sur le projet et son indexation (nom, durée, numéro de tâche, lien vers l'état d'avancement).

b.2.2) Requêtes

La première grande fonctionnalité offerte par Sémato est de permettre à l'utilisateur de procéder à des requêtes sur son corpus.

Il y a deux catégories principales de requêtes : repérage et analyse.

1)) Repérage

Les « requêtes en repérage » permettent de faire des recherches dans le corpus afin d'y trouver des éléments textuels spécifiques.

Une requête en repérage peut être faite selon trois modalités de base (dénommés « Ingrédients »).

- Les « Catégories de projet » regroupent les variables descriptives des données textuelles fournies au moment de l'indexation des textes (voir ci-dessus).

- Le repérage par « Thèmes » permet de retrouver les phrases et les textes catégorisés avec des thèmes (voir ci-après). Il suppose que des thèmes aient été introduit dans le projet manuellement ou automatiquement.

- Enfin, la modalité « Recherche textuelle » correspond à un moteur de recherche textuelle.

Le « Mode Et » permet de combiner plusieurs de ces ingrédients.

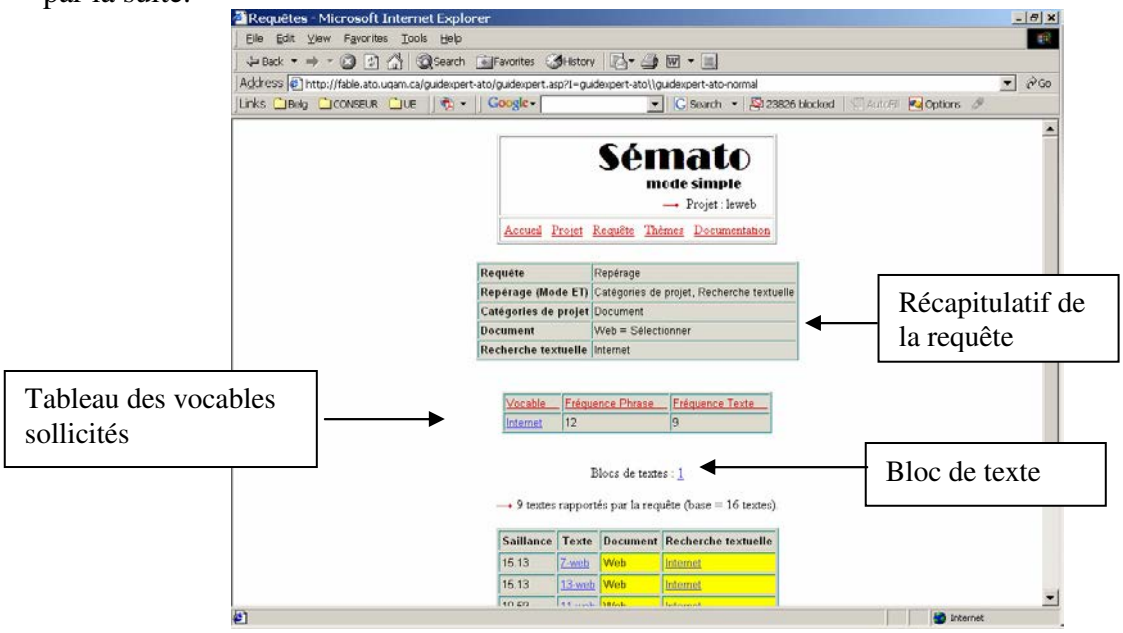
Chaque type de requête en repérage suppose différents choix dont un récapitulatif figure ci-après.

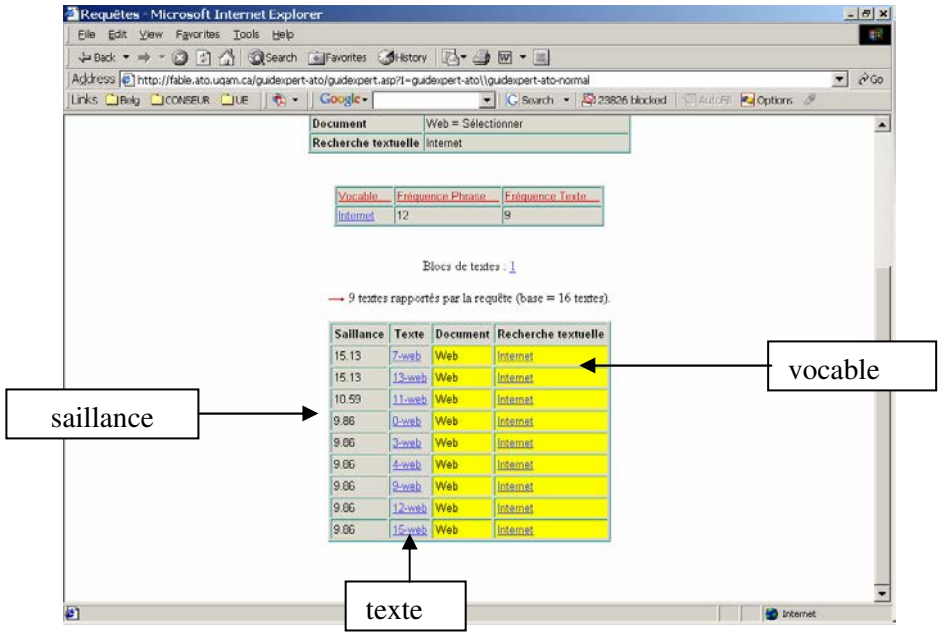
Une requête en repérage a été testée selon les critères suivants :
- modalités « Catégories de projets » et « Recherche textuelle »
- catégories de projets[157] : Document.
- textes sur lesquels on veut procéder à la requête : « Le Web »
- introduction de termes[158] pour la « Recherche textuelle » : Internet

Les deux écrans suivants montrent la fenêtre de résultat. En effet, celle-ci étant assez grande, elle a dû être découpée en deux.

[157] Une catégorie est automatiquement ajoutée aux « Catégories de projet » : la catégorie « Séquence ». Cette catégorie est de nature numérique, elle prend pour valeur un entier entre 0 et 9 (inclus). Cette valeur indique l'emplacement d'un texte (unité textuelle de Sémato) dans le fichier qui le contient.
[158] La case « Avec les champs sémantiques » permet de mettre en action les champs sémantiques pour fortifier la requête.
Le champ sémantique est une « liste de lemmes ou de synapsies associés à un lemme ».
Une synapsie est une expression plus ou moins figée construites sur un nom ou un verbe.
http://fable.ato.uqam.ca/guidexpert-ato///geadoc-vocabu.asp

Globalement, les fenêtres de résultat de Sémato se présentent toujours de la même manière. Nous allons donc décrire une seule fois celles-ci et ne présenter que les fenêtres différentes par la suite.





Le premier tableau donne un récapitulatif de la requête.

Le deuxième tableau correspond au « Tableau des vocables sollicités » dans la requête. Le nombre de phrases et de textes comprenant le vocable est affiché. Cliquer sur un vocable ouvre une fenêtre contenant tous les contextes de ce vocable.

Le lien « Bloc de texte » ouvre une fenêtre contenant les 50 premiers textes répondant à la requête.

Le dernier tableau présente les résultats.

La saillance donne le score relatif de chacun des textes rapportés par la requête. Son total est égal à 100.
La colonne « texte » indique les textes (unités textuelles) répondant à la requête.
La colonne « vocable » contient les termes entrés pour la recherche textuelle.

Il est possible d'afficher les résultats de différentes manières en cliquant sur les liens présents dans la page de résultats.

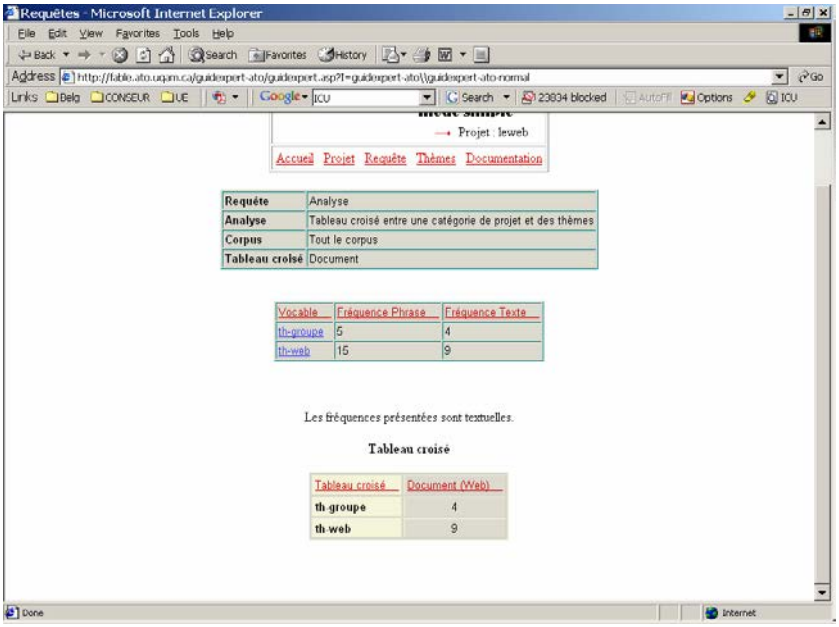
A chaque fois, des pages d'arrimages sont affichées (voir ci-après).

2)) Analyse

Les requêtes en analyse comprennent différentes catégories :
- Tableau croisé entre une catégorie de projet et des thèmes ;
- Cooccurrences des thèmes ;
- Réseaux de similitude ;
- Tableur catégories et thèmes.

1) L'analyse « Tableau croisé entre une catégorie de projet et des thèmes » permet de repérer les catégories de projets contenant un thème précis.

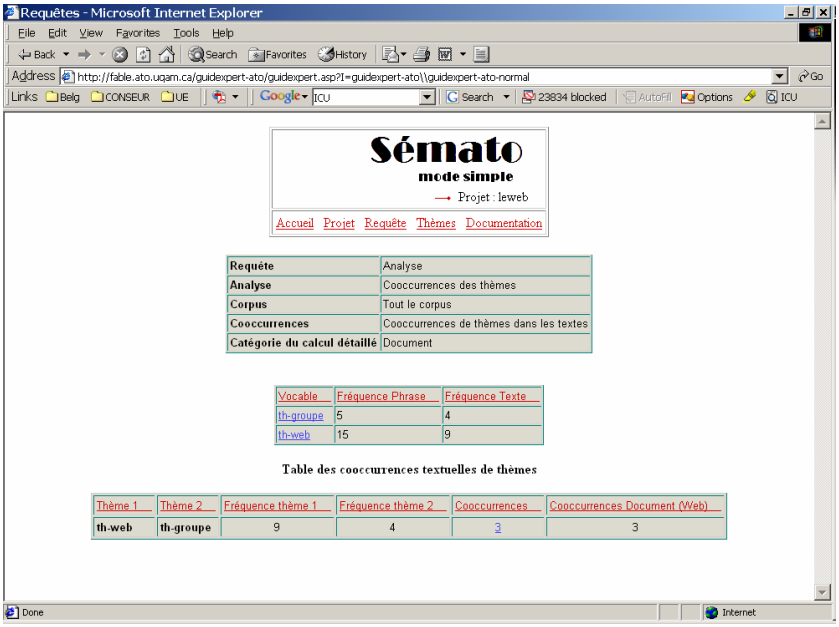
Selon son choix, l'utilisateur doit spécifier divers éléments permettant à l'analyse de se faire. Un récapitulatif figure ci-après.
La requête a donné le résultat suivant.



La page contient trois tableaux. Le premier présente un récapitulatif de la requête. Le second donne les fréquences globales en phrases et en textes pour chacun des thèmes retenus. Le troisième présente la matrice des fréquences textuelles des thèmes pour chacun des documents.

Le tableau 2 signifie que le thème « web » est présent dans 15 phrases et 9 textes.

2) La requête « Cooccurrences des thèmes » permet, comme son nom l'indique, de rechercher des cooccurrences entre les différents thèmes ajoutés au corpus. À nouveau, elle peut se faire sur tout le corpus ou une partie de celui-ci. La recherche peut se faire sur tous les thèmes ou seulement certains.

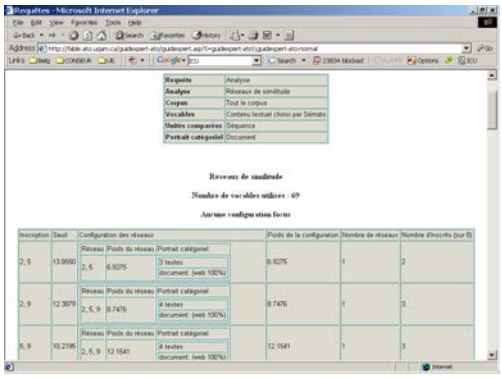


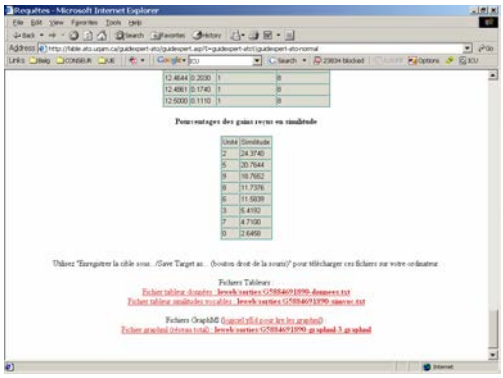
L'écran de résultat nous indique que les thèmes « web » et « groupe » sont cooccurrents dans trois textes du corpus.

3) Les « réseaux de similitude » permettent de voir émerger des groupes de ressemblance pour les valeurs d'une catégorie de projet donnée. Il s'agit de repérer des parties de texte présentant des similitudes.
Différents paramètres doivent être configurés. Il est notamment demandé à l'utilisateur de choisir entre une ressemblance fondée sur le « *contenu textuel choisi par Sémato* » ou une ressemblance fondée sur les « *propres thèmes* » de l'utilisateur.

Le « *contenu textuel choisi par Sémato* » signifie qu'on laisse à Sémato le soin de choisir les éléments de description linguistique qu'il juge les plus appropriés pour construire le modèle de ressemblance. Malheureusement, il n'a pas été possible d'identifier ces éléments.

L 'écran de résultat est composé de différents éléments :





Dans ces différents tableaux, Sémato fournit les différents éléments ayant permis son analyse ainsi que les résultats obtenus.

Ainsi, il donne le nombre de vocables (entités linguistiques) qu'il a utilisé dans la comparaison des textes. Il indique également s'il a trouvé une « configuration focus ». Il y a « configuration focus » lorsque des « îlots de ressemblance » ont pu être trouvés.

Pour trouver des îlots de ressemblance, les réseaux de similitude vont, en premier lieu, créer une liste de tous les couples d'objets à comparer et vont trouver, pour chaque couple le degré de leur ressemblance.

Sémato fait par la suite le total des degrés de ressemblance de tous les couples et construit une table où chaque couple est muni de son pourcentage de ressemblance (le rapport sur 100 de son propre degré sur le total).

Il s'agit ensuite de trouver un seuil en deçà duquel deux objets se dissemblent plus qu'ils ne se ressemblent afin d'identifier les ressemblances les plus fortes.
Une fois ce seuil identifié, le logiciel ne garde que les liens dont la valeur est au-delà de ce seuil. Tous ces liens vont former un réseau.

On obtient donc la configuration des réseaux construits par les liens les plus forts, c'est-à-dire la « configuration focus ».

Il est également possible de charger différents documents contenant des éléments ayant permis de calculer ces résultats.

4) Enfin, la requête « Tableur catégories et thèmes » permet de créer un fichier tableur de tous les textes du projet. Il y a dans ce tableur, autant de rangées que de textes et autant de colonnes que de catégories de projet et de thèmes sélectionnés.

Pour les thèmes, il est possible de choisir différents types de valeur numérique : 0 ou 1, fréquence absolue, fréquence relative.

Le résultat de cette requête donne un fichier .txt enregistrable sur le disque dur.

3)) Synthèse de l'outil requête

Afin d'aider le lecteur, une synthèse sous forme de table des matières de l'outil « Requête » est présentée ci-dessous. Elle permet de voir les choix à effectuer à chaque niveau.

Requête

1. repérage
    a. catégories de projet
        i. choix catégorie
            1. choix numérique si séquence
    b. thèmes
        i. choix de thème(s)
    c. recherche textuelle
        i. introduction des termes
2. analyse
    a. Tableau croisé entre une catégorie de projet et des thèmes
        i. Tout le corpus
            1. choix catégorie : si document
                a. Sélection manuelle des thèmes
                    i. Sélection des thèmes
                b. Sélection de tous les thèmes à la fois
            2. choix catégorie Séquence
        ii. Section du corpus
            1. Catégories de projets : choix
                a. Document
                    i. Choix du document
                b. Séquence
                    i. Choix numérique pour la séquence
            2. Thèmes
                a. Choix des thèmes
    b. Cooccurrences des thèmes
        i. Tout le corpus
            1. Cooccurrences de thèmes dans les textes
                a. Sélection manuelle des thèmes
                    i. Sélection des thèmes
                        1. choix catégorie
                b. Sélection de tous les thèmes à la fois
            2. Cooccurrences de thèmes dans les phrases
                a. Sélection manuelle des thèmes
                    i. Sélection des thèmes
                        1. choix catégorie
                b. Sélection de tous les thèmes à la fois
        ii. Section du corpus
            1. Catégories de projets
                a. Choix catégorie
            2. Thèmes
    c. Réseaux de similitude
        i. Tout le corpus
            1. Contenu textuel choisi par Sémato
                a. Document (unité comparée)
                b. Séquence (unité comparée)
                    i. Choix catégories
            2. Les thèmes sélectionnés au prochain dialogue
                a. Sélection manuelle des thèmes
                    i. Sélection thèmes
                        1. choix catégories
                            a. autre choix catégorie
                b. Sélection de tous les thèmes à la fois
        ii. Section du corpus
            1. Catégories de projets
                a. Choix catégorie
            2. Thèmes
    d. Tableur catégories et thèmes
        i. Tout le corpus
            1. Choix catégorie
                a. 0 ou 1
                    i. Sélection manuelle des thèmes
                        1. choix thème
                    ii. Sélection de tous les thèmes à la fois
                b. fréquence absolue
                c. fréquence relative
                    i. choix d'un nombre
        ii. Section du corpus
            1. Catégories de projet
                a. Choix catégorie
            2. Thèmes

b.2.3) Outil « Thème »

La deuxième grande fonctionnalité d'analyse offerte par Sémato est l'outil « Thèmes » qui permet à l'utilisateur de thématiser son corpus manuellement ou automatiquement.

Sémato offre deux outils. L'assistant scripteur de thème qui aide l'utilisateur à créer ses thèmes manuellement, et l'outil GHT qui génère des thèmes automatiquement.

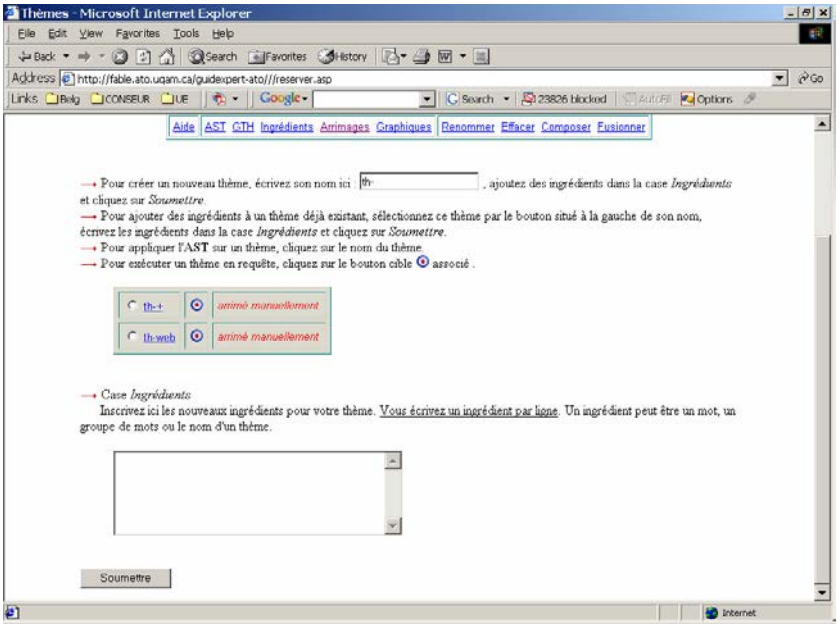
1)) AST : l'assistant scripteur de thèmes

L'Assistant Scripteur de Thèmes (AST) assiste l'utilisateur dans l'écriture des thèmes de son corpus.

Un thème est caractérisé par des ingrédients. Un ingrédient peut être un mot, un groupe de mots ou le nom d'un autre thème.

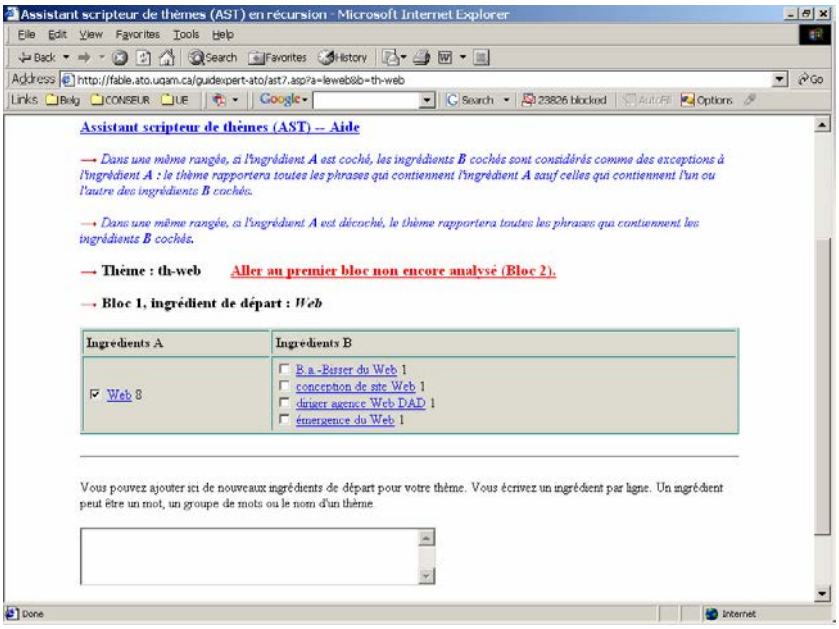
L'AST a pour fonction d'analyser les ingrédients d'un thème et de proposer d'autres ingrédients associés sémantiquement aux premiers.

Voici l'écran qui se présente à l'utilisateur lors de l'accès à la fonctionnalité « Thèmes ».



Ainsi, il est possible de créer de nouveaux thèmes, d'affiner les thèmes existants avec de nouveaux ingrédients, d'appliquer l'AST sur les thèmes existants afin d'obtenir de nouveaux ingrédients et d'appliquer un thème au corpus.

L'application de l'AST au thème « th-web » obtenue en cliquant sur ce thème donne l'écran suivant.

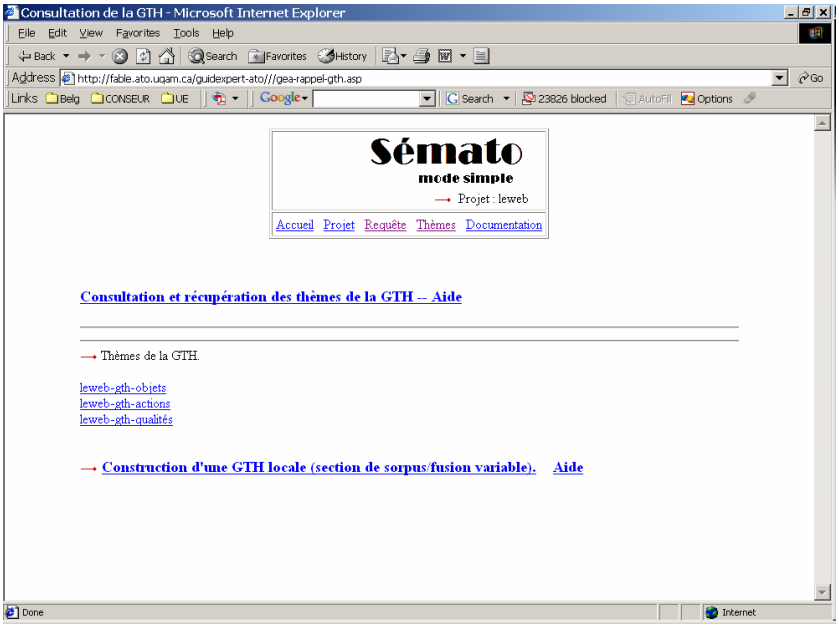


Dans le bloc, l'AST propose d'autres ingrédients. Pour chacun de ces éléments, il propose aussi des synapsies ou petites unités de contexte. Les synapsies sont présentées dans la colonne « Ingrédients B ». Ceux-ci permettent de circonscrire rapidement le sens d'un Ingrédient A.

2)) Génération de thèmes – GHT

Le deuxième outil de thématisation est GHT qui permet de générer automatiquement des thèmes.

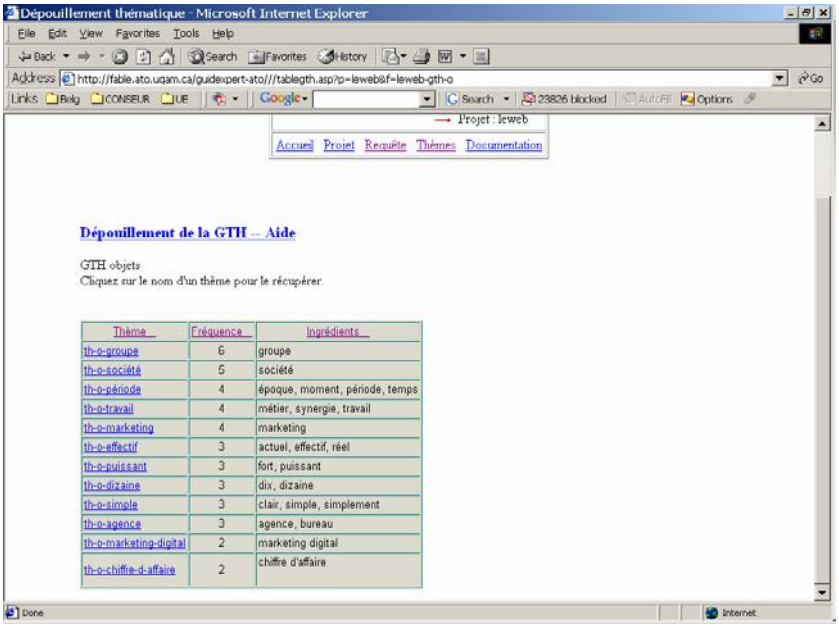
La GTH est commandée automatiquement par la fonction d'indexation.
A partir de la page « Thème », il suffit de cliquer sur GHT.



Trois tables de thèmes sont construites par la GHT : Objets, Actions et Qualités.

Les thèmes *Objets* rassemblent ce dont on parle, les *Actions* rassemblent ce que l'on fait et les *Qualités* rassemblent des qualifications.

L'écran suivant montre les « Objets ».



Les thèmes générés par la GTH peuvent être modifiés par l'utilisateur.

À cet effet, chaque thème de la table contient un lien qui permet de le récupérer et de le manipuler.

La soumission du thème aura pour effet que celui-ci pourra être utilisé via les outils « Requêtes » et AST.

3)) Outil « Ingrédients »

L'outil « Ingrédients » permet de créer des thèmes depuis une liste d'ingrédients et en indiquant une fréquence minimale.

4)) Autres outils

La page « Thèmes » offre également des outils permettant de nommer, effacer et fusionner des thèmes. Dans le cadre de la fusion, le nouveau thème contiendra tous les ingrédients des thèmes fusionnés. Les thèmes fusionnés seront détruits. Tous les ajouts et les retraits d'arrimage associés aux thèmes fusionnés seront, à la suite de cette opération, associés au nouveau thème résultat de la fusion.

b.2.4) Pages d'arrimage

Les pages d'arrimage donnent accès aux textes et phrases du projet, permettent d'ajouter ou de retirer des thèmes et d'associer des mémos analytiques.

Les « thèmes » ont pour fonction de catégoriser les phrases et textes du corpus. Cette catégorisation peut se faire de manière automatique ou manuelle (voir ci-avant).

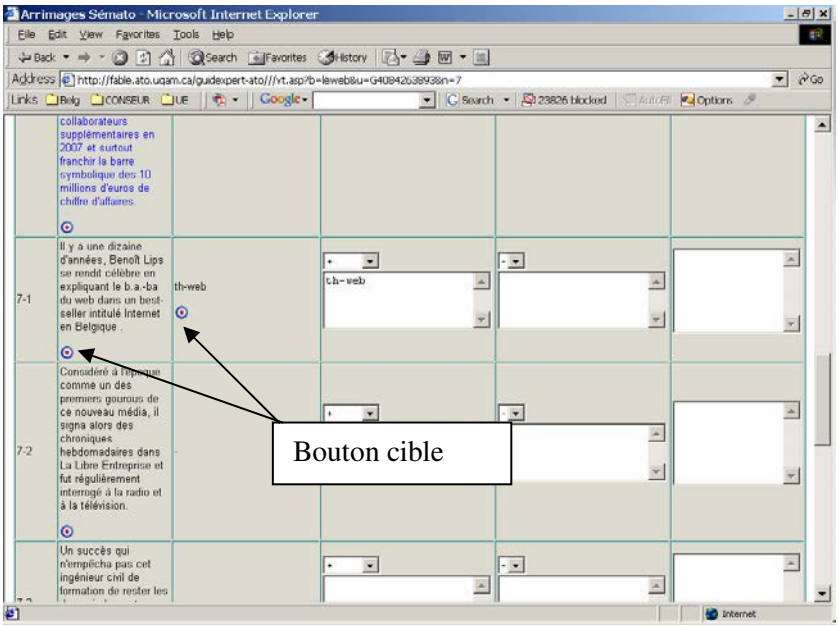
Les « mémos analytiques » sont des notes de travail associées aux phrases ou textes du corpus.

Ces pages d'arrimage sont accessibles de deux manières : via les résultats des requêtes ou via la page « Thèmes ».

Il y a trois types de page d'arrimage : arrimage sur un ensemble de textes, arrimage sur un seul texte ou arrimage sur un ensemble de phrases.

Une fois qu'un thème est inséré, il est disponible dans le menu des Ajouts d'arrimage thématique.

De plus, chaque unité de la page d'arrimage contient un bouton cible permettant de rechercher touts les textes/phrases ayant les mêmes thèmes ou ressemblant le plus à la phrase cible.



b.3) Affichage et autres sorties

b.3.1) Graphiques

Sémato permet de construire une représentation graphique du contenu des thèmes.
Ces graphiques sont lisibles via le logiciel « yEd Graph Editor » téléchargeable sur le site de Sémato.

Trois types de représentation sont possibles.

La première option permet de représenter tous les ingrédients de tous les thèmes.
La seconde option restreint le graphique aux seuls thèmes composés d'autres thèmes.
La troisième option permet de dessiner le graphique du réseau des champs sémantiques des ingrédients de thèmes sélectionnés.

Sémato permet également de choisir un niveau de profondeur pour le réseau (automatique ou de 0 à 4). Le niveau 0 correspond au champ immédiat des ingrédients du thème. Le niveau 4 compte 4 liaisons entre les ingrédients immédiats du thème et les plus éloignés des nœuds du réseau qui sera dessiné.

b.3.2) Fichiers txt.

Lors de certaines opérations, des fichiers .txt sont produits et peuvent être sauvegardés sur l'ordinateur de l'utilisateur.

b.3.3) Sauvegarde des données du projet

Le lien « sauvegarde des données du projet » de la page « Projet » donne accès à une page contenant la liste des ajouts d'arrimage et des retraits d'arrimage thématique, des mémos analytiques ainsi que des thèmes du projet.

c) Résumé

Le logiciel Sémato est un outil d'analyse de texte en ligne. Il présuppose le téléchargement des textes à analyser sur le serveur du site.

Il peut être classé parmi les logiciels permettant une analyse socio-sémantique. En effet, certaines opérations effectuées se basent sur des théories préexistantes (notamment le « configuration focus » et la génération automatique de thèmes).
Sémato apporte également un outil pour l'analyse par réseau de mots associés (recherche des cooccurrences).

Le transfert du texte procède à une « indexation » de celui-ci, c'est-à-dire un découpage en unités textuelles, celles-ci correspondant aux retours chariot.

En ce qui concerne les fonctionnalités d'analyse,

- Une des fonctionnalités principales de Sémato est de permettre d'effectuer des recherches de divers types dans le corpus selon différents critères : recherche de termes, recherche de cooccurrences, recherche de thèmes, recherche de configuration focus, etc.
- Le deuxième outil principal de Sémato est son analyseur thématique permettant de créer ses propres thèmes, mais également de générer ceux-ci automatiquement.

Sémato offre également un outil « Pages d'arrimages » permettant d'attacher des thèmes ou mémos analytique aux unités textuelles et phrases du corpus.

Dans certains cas, des fichiers de résultats sauvegardables sont produits.

2.5. TROPES

a) Généralités

Le logiciel Tropes est le seul représentant de l'analyse cognitivo-discursive.

Pour réaliser l'analyse de ce logiciel, nous nous sommes basés sur sa version démonstration téléchargeable sur le site d'Acetic et sur le manuel fourni avec cette version « Tropes® *Version 7.0* Manuel de référence ».

Lors du lancement du logiciel, Tropes informe l'utilisateur sur la manière de commencer une analyse. Il lui indique la manière de lancer celle-ci et où trouver les principales fonctions. Il rappelle également que la version d'évaluation est limitée quant à la quantité de texte pouvant être analysée simultanément.

b) Fonctionnalités

b.1) Formats d'entrée

Le logiciel Tropes accepte différents formats de fichier en entrée : ANSI, html, Word, RTF, Acrobat, etc., parfois moyennant des conversions.

Le texte analysé peut contenir des signes de ponctuation, des majuscules et des caractères spéciaux (des parenthèses, des nombres, des pourcentages, etc.). Seules les lettres de l'alphabet et les caractères de ponctuation seront utilisés durant l'analyse.

Il est parfois intéressant de considérer une suite de mots comme un mot unique (ex. : « orienté objet »). Dans ce cas, il faut lier les mots par le caractère de soulignement « _ » (ex. : « orienté_objet »).

b.2) Analyse du corpus et fonctions

Le lancement de l'analyse du corpus se fait automatiquement dès la sélection et l'ouverture d'un fichier contenant le texte à analyser.

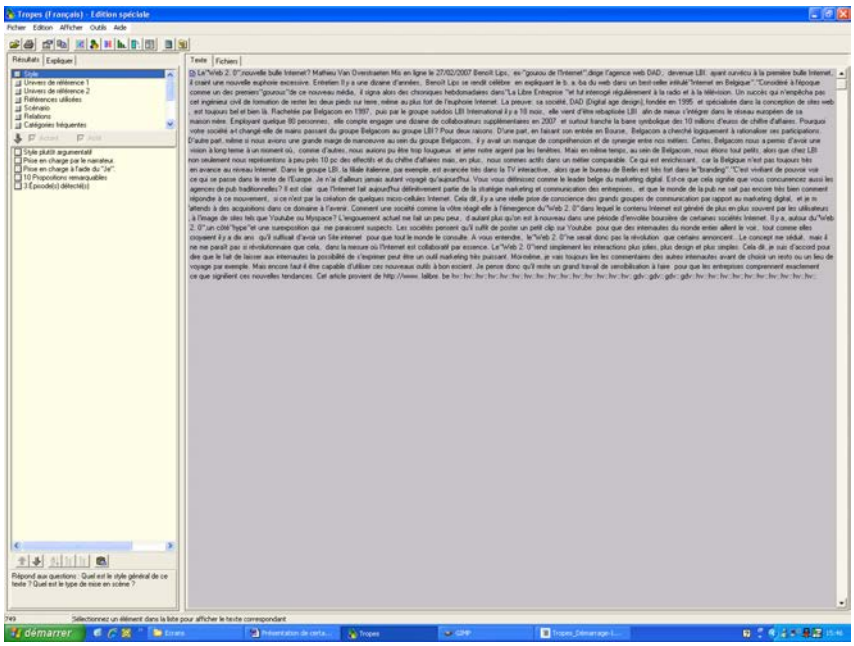
Pour effectuer cette analyse, Tropes procède en six étapes :
1. découpage des phrases et des propositions,
2. levée d'ambiguïté des mots du texte,
3. identification des *classes d'équivalents*,
4. statistiques, détection des *rafales* et des *épisodes*,
5. détection des *propositions remarquables*,
6. mise en forme et affichage du résultat.

Le découpage en propositions se fait sur base d'une analyse de la ponctuation et d'une analyse syntaxique.

Au niveau statistique, plusieurs analyses sont faites :
- des statistiques sur la fréquence globale d'apparition des grandes *catégories de mots*, et de leurs sous-catégories,
- des statistiques sur la cooccurrence et le taux de liaison des *classes d'équivalents* et des *catégories de mots*,
- une analyse des mots arrivant en *rafales* et une analyse des *rafales* délimitant les *épisodes*,
- une analyse cognitivo-discursive (ACD) permettant de détecter les *propositions remarquables*.

Les « Catégories de mots fréquentes » et le « Style général du texte » (voir ci-après) sont obtenus en comparant la répartition des fréquences d'apparition des catégories observées dans le texte avec des normes de production langagière. Ces normes ont été élaborées en étudiant un grand nombre de textes différents et sont stockées dans des tables internes au logiciel.

Immédiatement, les résultats de l'analyse sont affichés.

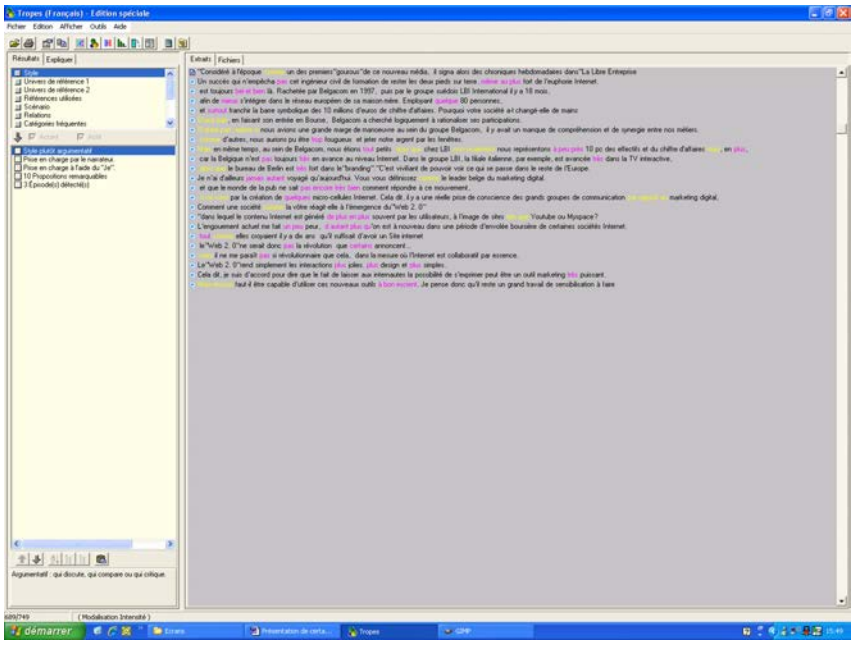


La partie gauche, consacrée aux différents types d'analyse (fonctionnalités) procédés par Tropes, est divisée en deux parties. La partie supérieure reprend les catégories d'analyses principales effectuées par Tropes, tandis que la partie inférieure présente les subdivisions de ces catégories principales.

b.2.1) Fonction « Style »

La fonctionnalité style est décomposée en différentes sous-fonctions :

- Style ;
- Mise en scène verbale (représentée ici par « Prise en charge par le narrateur » et « Prise en charge à l'aide du « Je » ») ;
- Propositions remarquables ;
- Episodes détectés.



En ce qui concerne le « Style », Tropes détecte celui-ci en fonction des indicateurs statistiques récupérés au cours de l'analyse. Il importe donc que le texte soit suffisamment long pour que ces résultats soient significatifs.

Quatre styles sont possibles :

Style	**Description**
Argumentatif	Le sujet s'engage, argumente, explique ou critique pour essayer de persuader l'interlocuteur.
Narratif	Un narrateur expose une succession d'événements, qui se déroulent à un moment donné, en un certain lieu.
Enonciatif	Le locuteur et l'interlocuteur établissent un rapport d'influence, révèlent leurs points de vue.
Descriptif	Un narrateur décrit, identifie ou classifie quelque chose ou quelqu'un.

Dans le cadre de l'exemple, le logiciel Tropes a conclu que le style était argumentatif. La partie droite de l'écran reprend les phrases du texte qui l'ont conduit à ce résultat.

La « mise en scène verbale », effectuées également sur base des résultats statistiques, comprend quatre types possibles :

Mise en scène	**Expression**
Dynamique, action	des verbes d'action
Ancrée dans le réel	des verbes de la famille d'être et avoir
Prise en charge par le narrateur	des verbes qui permettent de réaliser une déclaration sur un état, une action, ...
Prise en charge à l'aide du « Je »	de nombreux pronoms à la première personne du singulier (« je », « moi », « me », ...)

Dans le cadre de l'exemple, deux types de mise en scène ont été retenus : « Prise en charge par le narrateur » et « Prise en charge à l'aide du « Je » ». Le fait de cliquer sur celles-ci dans la partie gauche de l'écran fait apparaître les phrases clés de ce résultat dans la partie droite.

Les « Propositions remarquables » sont obtenues par contraction du texte. Ce sont « *des propositions qui introduisent des thèmes ou des personnages principaux, qui expriment des événements nécessaires à la progression de l'histoire (attributions causales, des conséquences, des résultats, des buts)* ».[159]

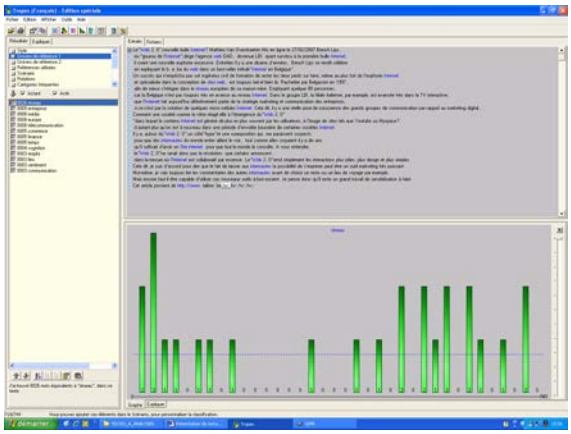
Tropes effectue une analyse cognitivo-discursive (ACD) afin d'extraire ces propositions. Chaque proposition du texte se voit attribuer un score calculé en fonction de son poids relatif, de son ordre d'arrivée et de son rôle argumentatif. Les propositions sont ensuite filtrées en fonction de leur score.

De l'aveu des concepteurs, les propositions remarquables n'ont de sens que par rapport à un texte monolithique, structuré et pas trop long. Ainsi, dans le cadre de textes trop long, elles ne seront pas affichées.

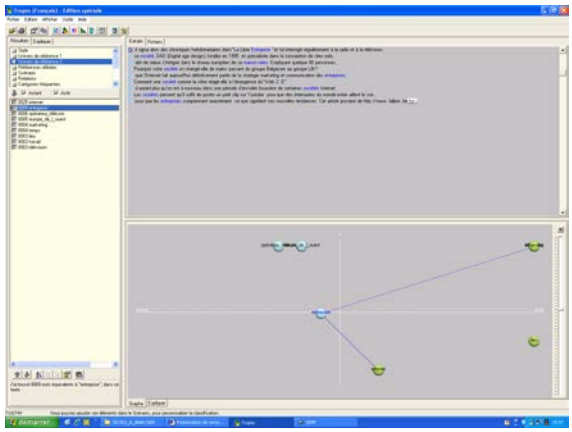
b.2.2) Univers de référence

Cette fonction affiche par fréquence décroissante, les « Univers de référence » du texte. Dans la partie gauche inférieure, chaque ligne comporte un « Univers », précédé d'un compteur indiquant le nombre de mots (occurrences) qu'il contient. Seuls les « Univers » significatifs sont affichés.

Univers de référence 1



Univers de référence 2



Les univers de référence regroupent, dans des classes d'équivalents, les principaux substantifs du texte.

La sélection de ces univers se base sur un dictionnaire sémantique ne contenant pas tous les termes du français. Seuls les substantifs les plus significatifs et certains noms propres apparaissent.

La sélection d'un univers permet de voir s'afficher dans la partie droite tous les mots du texte analysé qui en font partie.

[159] cfr Manuel.

Différence Univers de référence 1 et Univers de référence 2

Les « classes d'équivalents » regroupent les références (noms communs ou noms propres) qui apparaissent fréquemment dans le texte et qui possèdent une signification voisine.

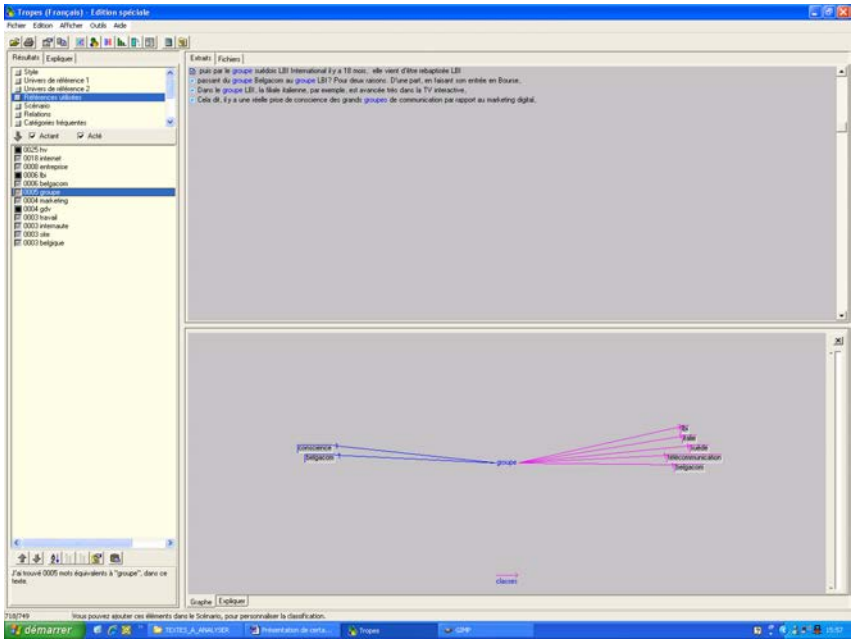
Les « Univers de référence » regroupent les mots contenus dans les classes d'équivalents. Au niveau le plus bas se situent les « Références utilisées », elles-mêmes regroupées de façon plus large dans les « Univers de référence 2 », qui à leur tour sont regroupés dans les « Univers de référence 1 ».

Exemple de regroupement repris du manuel :

Univers 1	Univers 2	Classes	Mots
politique	doctrine politique	communisme	communisme
politique	doctrine politique	communisme	marxisme
politique	doctrine politique	libéralisme	capitalisme
politique	doctrine politique	libéralisme	libéralisme
politique	homme politique	chef d'état	chef d'état
politique	homme politique	chef d'état	président de la république
politique	homme politique	ministre	garde des sceaux
politique	homme politique	ministre	ministre
politique	homme politique	parlementaire	député
politique	homme politique	parlementaire	sénateur
politique	instance politique	gouvernement	gouvernement

b.2.3) Références utilisées

Cette fonction affiche, regroupés par « classes d'équivalents » (noms communs et noms propres ayant un sens voisin) et triés par fréquence décroissante, les substantifs (*références*) utilisés dans le texte. À nouveau, seules les références significatives sont affichées.



b.2.4) Scénarios

L'outil Scénario permet d'enrichir et de filtrer les « classes d'équivalents », c'est-à-dire les regroupements de termes proches sémantiquement.

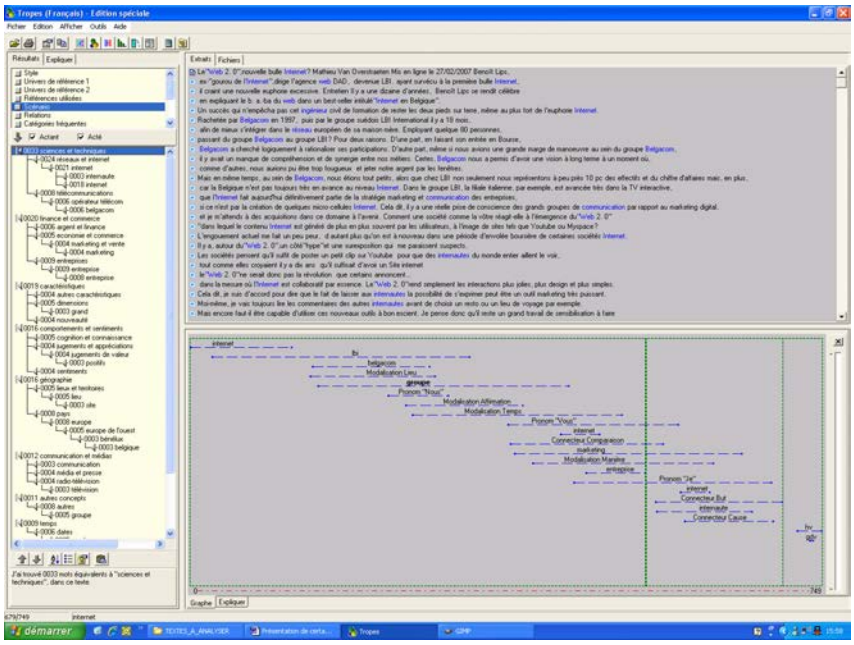
Cet outil permet de définir ses propres classifications, ou de modifier ou restructurer les dictionnaires du logiciel.

Tropes est livré avec différents scénarios préétablis :

- global (*Scénario Concept Fr V7* qui regroupe les références par grands thèmes, à la manière d'un thésaurus généraliste grand public) ;
- détaillé (*Scénario Concept Fr V7 détaillés* qui regroupe les références dans un plus grand nombre de thèmes, à la manière d'une encyclopédie) ;
- très spécialisée (autres Scénarios, ou ceux disponibles sur demande).

Un « Scénario » est constitué d'un certain nombre de groupes sémantiques, c'est-à-dire de regroupements de mots et/ou de classes d'équivalents, qui peuvent être hiérarchisés sur neuf niveaux de profondeur.

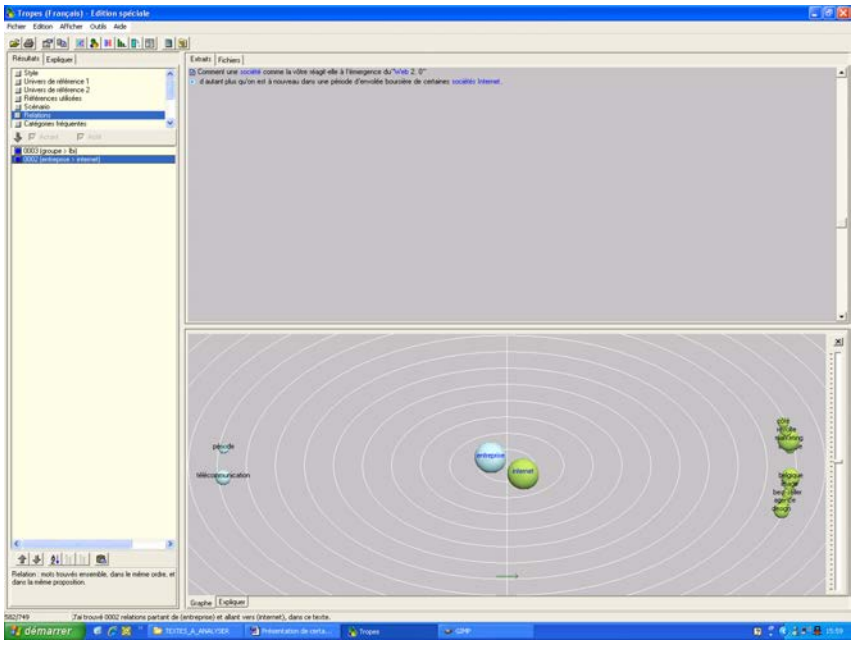
Il est possible de créer ses propres scénarios et de les utiliser sur son corpus. On peut y mettre des mots, des classes d'équivalents et des lemmes de verbes ou d'adjectifs.



b.2.5) Relations

La fonction « Relations » permet d'afficher, triées par fréquence décroissante, les relations de cooccurrence entre les classes d'équivalents, ainsi que la fréquence des ces relations.

Les relations indiquent quelles classes d'équivalents sont fréquemment reliées dans le texte analysé. Ces relations sont orientées suivant l'ordre d'apparition des mots qui les composent (généralement en allant des actants vers les actés, c'est-à-dire dans le sens de lecture).

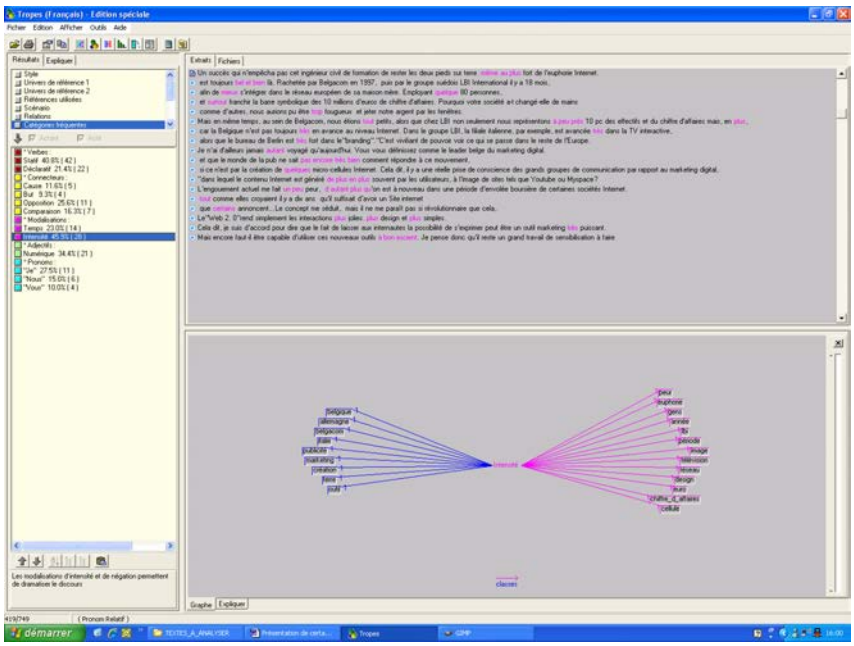


b.2.6) Catégories fréquentes

Cette fonction affiche les catégories de mots les plus significatives : verbes, connecteurs, adjectifs, modalisations, pronoms.

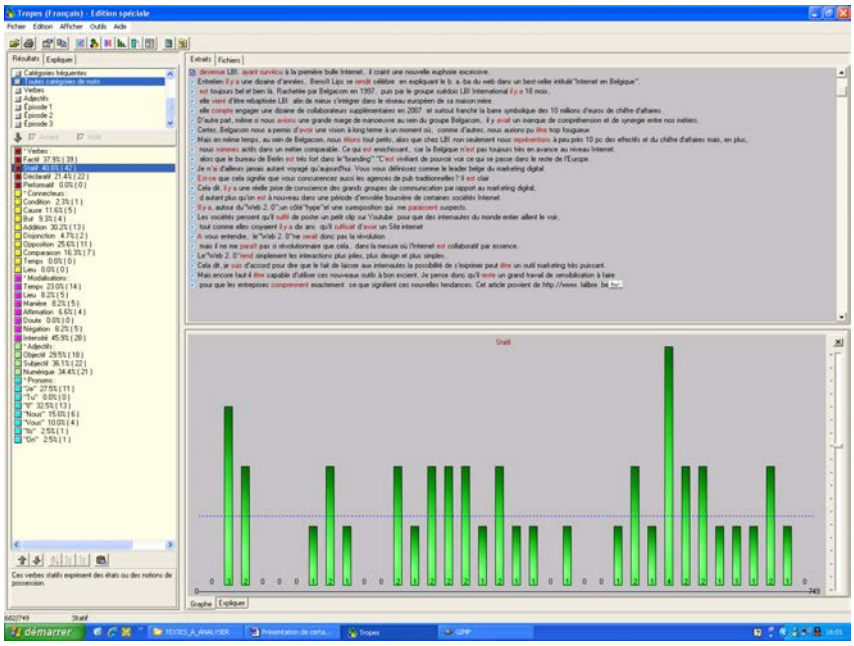
Une « catégorie de mots » est considérée comme significative lorsque sa fréquence d'apparition est nettement supérieure à la moyenne. Ces résultats sont construits en comparant les statistiques du texte analysé avec des tables internes au logiciel.

Les verbes et adjectifs sont réduits à leur forme canonique (lemme).



b.2.7) Toutes catégories de mots

Cette fonction affiche toutes les « catégories (et sous-catégories) de mots » du texte analysé. Chaque ligne dans la partie gauche inférieure de l'écran comprend une catégorie, sa répartition dans la sous-catégorie concernée (pourcentage) et le nombre d'occurrences trouvées.



Dans l'exemple, les verbes sont en rouge, les connecteurs en jaune, les modalisations en fuchsia, les adjectifs en vert et les pronoms en bleu.

Signalons que les graphes en aires ne sont pas disponibles pour cette fonction.

b.2.8) Épisodes

La fonction « Episodes » (accessibles directement ou via la fonction « Style ») permet d'étudier la chronologie du discours. Celle-ci se base sur deux notions : les rafales et les épisodes.

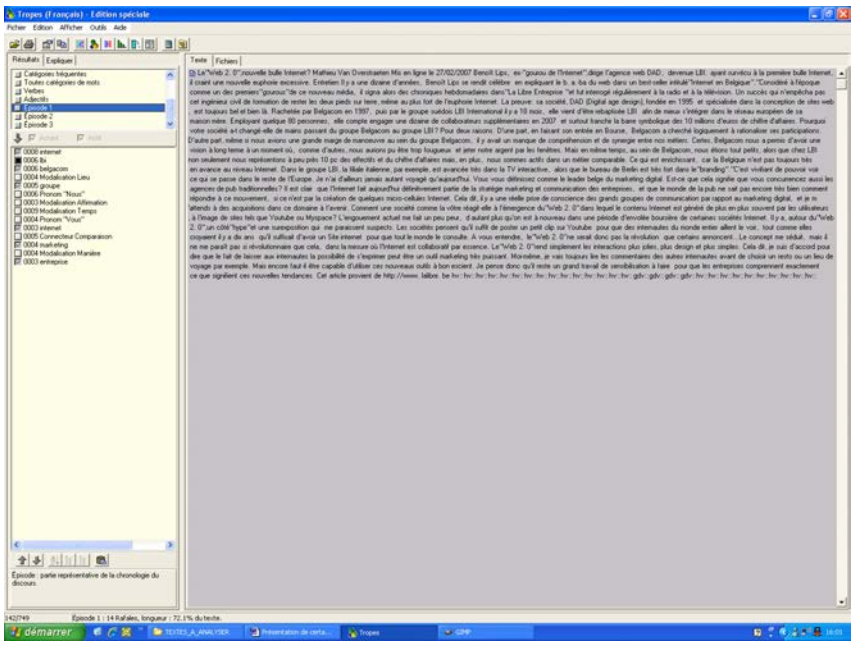
Une « Rafale » regroupe des occurrences de mots (contenus dans une classe d'équivalents ou une catégorie de mots) ayant tendance à arriver avec une concentration remarquable dans une partie limitée du texte (quelque part au début, au milieu ou à la fin du texte, mais jamais de façon uniforme sur l'intégralité de celui-ci).

Un « Episode » correspond à une partie du texte dans lequel un certain nombre de rafales se sont formées et terminées.

Les épisodes sont affichés les uns à la suite des autres, et numérotés en fonction de leur ordre d'arrivée dans le texte.

Dans chaque épisode, les rafales sont triées en fonction de leur adresse (moyenne de la position des mots) et préfixées par la fréquence d'occurrence des mots qui la composent.

Selon les concepteurs, ces éléments n'ont de sens que pour les textes de taille raisonnable et ne sont donc pas affichés pour les textes trop longs.



b.2.9) Fonctionnalités supplémentaires

1)) Analyse des acteurs

Dans le cadre des fonctions « Univers de référence », « Références » et « Groupes du scénario », il est possible d'effectuer une analyse des « acteurs » en cochant les cases correspondantes dans la partie gauche de l'écran.

Cet outil permet d'analyser ceux-ci en distinguant les termes placés en position :
- d'« actant », c'est-à-dire avant le verbe (et souvent sujet de ce dernier),
- d'« acté », c'est-à-dire après le verbe (et rarement sujet de ce dernier).

Selon les concepteurs, l'identification des actants et des actés constitue une des étapes fondamentales de l'analyse de texte. En effet, lorsqu'un univers de référence significatif (ou une référence utilisée) se trouve fortement placé en position d'actant (taux supérieur à 60 %) on peut généralement considérer qu'il effectue l'action. Dans le cas contraire, lorsqu'un univers (ou une référence utilisée) significatif se trouve fortement placé en position d'acté, on peut généralement considérer qu'il subit l'action.

2)) Recherche de termes

Tropes permet d'effectuer des recherches dans le corpus. Deux options sont possibles : la recherche du mot dans le texte, ou dans les classes d'équivalents ou groupes sémantiques du scénario.

3)) Délimiteur

Il est parfois intéressant de diviser un texte en plusieurs parties. Par exemple, si un texte reprend une discussion entre plusieurs personnes, on veut pouvoir isoler les interventions de chacun. De même, si on analyse un questionnaire, il est intéressant de séparer chaque question.

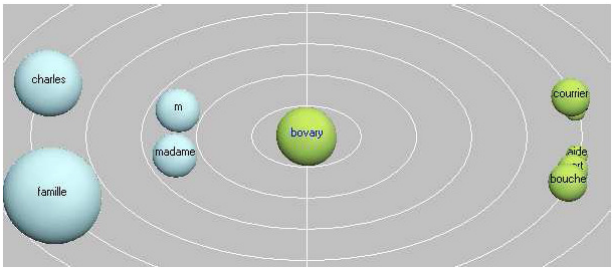
L'outil « Délimiteur », uniquement dans la version payante, permet cela. Sur base d'une préparation préalable du texte (introduction de codes spécifiques), il effectue une segmentation automatique de ce dernier.

b.3) Affichages et autres sorties

b.3.1) Affichages

Les différentes captures d'écran insérées ci-dessus ont déjà permis de voir que différents types d'affichage graphique étaient possibles : modes « étoilé », « aires », « répartition », « épisodes », « acteurs ».

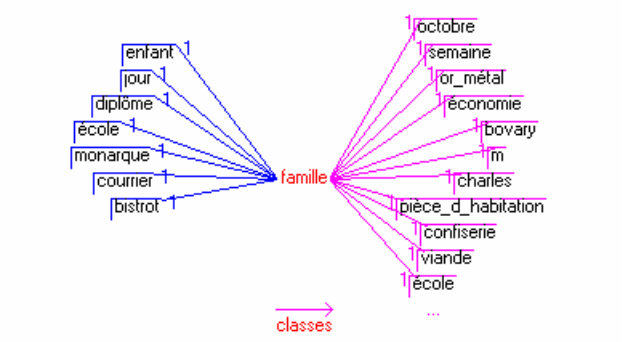
1)) Mode « aires » [160]



Dans le cadre des classes d'équivalents, la taille de chaque sphère est proportionnelle au nombre de mots qu'elle contient. La distance entre la classe centrale et les autres classes est proportionnelle au nombre de relations qui les lient.

Le recouvrement de deux sphères n'a pas de signification.
Le graphe en aire ne peut pas être affiché sur les méta-catégories de mots.
Il est possible d'augmenter le nombre de sphères affichées.

2)) Mode « étoilé » [161]

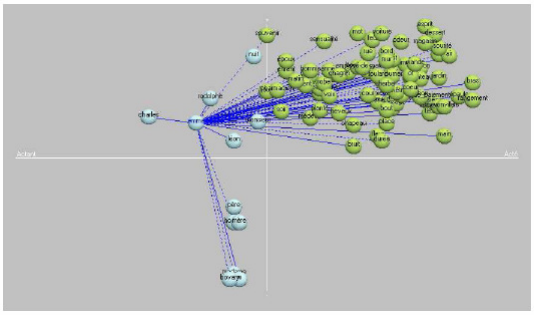


[160] Exemple provenant du manuel.
[161] Exemple provenant du manuel.

Ce type de graphe affiche les relations entre classes d'équivalents, ou entre une catégorie de mots et des classes d'équivalents.

Les nombres qui apparaissent sur le graphe indiquent la quantité de *relations* (fréquence de cooccurrence) existant entre les *classes d'équivalents*.

3)) Mode « acteurs »[162]



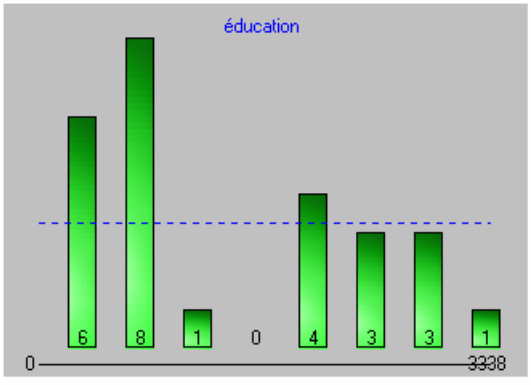
Le graphe des acteurs représente la concentration de relations entre les principaux acteurs (actants/actés) sur la totalité du texte. Il permet de faire une comparaison visuelle du "poids" des relations entre les principales références (ou bien entre les groupes du Scénario).

L'axe des X (horizontal) indique le taux actant/acté (de gauche à droite).
L'axe des Y (vertical) indique la concentration de relations pour chaque référence affichée (forte en haut du graphe, et faible en bas).

La concentration de relations est calculée en divisant, pour chaque référence, le nombre total de relations par le nombre de relations différentes.

Les traits indiquent les relations entre la référence sélectionnée et les autres références affichées. Un trait en pointillé indique une relation peu fréquente. Un trait plein indique une relation fréquente.

4)) Mode « répartition » [163]



[162] Exemple provenant du manuel.
[163] Exemple provenant du manuel.

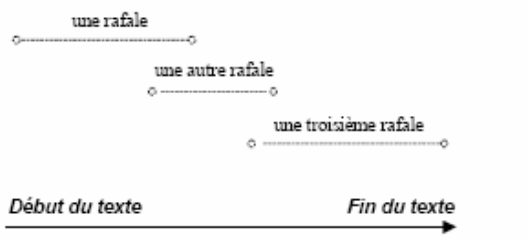
Les graphes par répartition permettent d'afficher un histogramme de répartition (chronologique) d'une classe d'équivalents, d'une mise en relation (de deux classes d'équivalents), ou d'une catégorie de mots.

Pour construire ce graphe, le texte est divisé en secteurs contenant un nombre égal de mots. Puis, le logiciel calcule la fréquence d'apparition de la classe d'équivalents ou de la catégorie de mots sélectionnée à l'intérieur de chaque secteur.
Les barres de l'histogramme affichent chaque secteur dans l'ordre chronologique, de gauche (début du texte), à droite (fin du texte). La ligne en pointillés indique la taille moyenne des barres de l'histogramme.

Le nombre de barres de l'histogramme et le nombre de mots contenus dans chaque barre sont déterminés automatiquement par le logiciel en fonction du nombre total de mots du texte et de la taille de la fenêtre principale.

Lorsque le graphe de répartition porte sur une relation, l'histogramme représente les fréquences d'occurrence cumulées des classes contenues dans la relation.

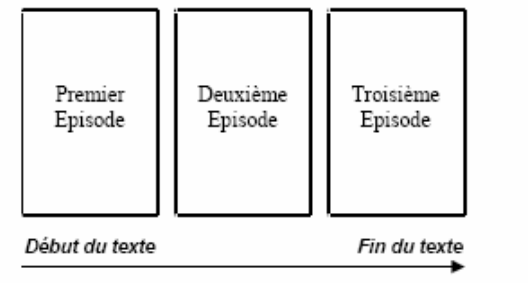
5)) Mode « épisodes »



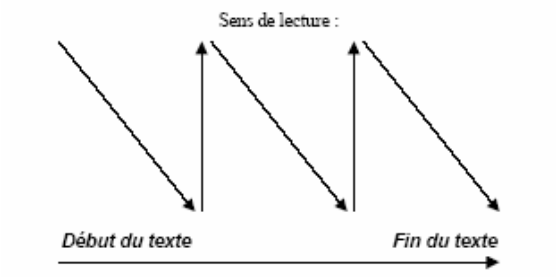
Ce mode permet de visualiser les épisodes et leurs rafales.

Chaque rafale est affichée sous la forme d'une ligne pointillée horizontale indiquant son étendue (longueur de la rafale) et sa position par rapport au début du texte.

Les épisodes sont affichés sur le même graphe, de gauche à droite, dans l'ordre chronologique, sous la forme de grands cadres en pointillés.



Les rafales sont affichées de haut en bas et de gauche à droite, en fonction de leur ordre d'arrivée dans le texte. Lorsque le bas de l'écran est atteint, le logiciel effectue un balayage en zigzag, pour afficher le plus d'information possible.



b.3.2) Impression

Il est possible d'imprimer les résultats et les graphes.[164]

b.3.3) Exportation

Il est possible de copier les graphes pour les placer dans le Presse-papiers Windows afin de les utiliser dans un traitement de texte compatible.

Cette fonction est également disponible pour les parties de texte affichées dans la partie droite supérieure et pour les résultats.

b.3.4) Génération de rapports

Un générateur d'état permet de construire automatiquement des rapports (contenant des tableaux statistiques et des graphiques) dans Microsoft Excel®, en utilisant les résultats obtenus avec le Scénario.

b.4) Présence d'un dictionnaire

Les analyses effectuées par Tropes se basent sur un dictionnaire incorporé dans le logiciel.

Quelques remarques doivent être faites sur celui-ci. Celles-ci sont reprises principalement du manuel.

Le logiciel génère automatiquement des classes d'équivalents pour tous les mots non référencés dans ses dictionnaires. Ces « classes générées » contiennent aussi des mots identifiés par le logiciel mais qui ne peuvent pas être regroupés par équivalents (par exemple, « *Charles* »).

Ces classes générées sont visibles uniquement dans les références utilisées, et précédées par un carré noir, alors que les classes d'équivalents sont précédées par un carré bleu.

Il est possible d'intégrer ces classes générées aux classes d'équivalents existantes et donc de créer sa propre classification en utilisant l'outil « Scénarios ».

[164] L'impression des **classes d'équivalents** et des **Scénarios** font apparaître un taux d'utilisation, exprimé en pourcentage, qui correspond au nombre de mots de chaque classe divisé par le nombre total de mots du texte.
L'impression des **relations** fait apparaître une information supplémentaire : le taux de liaison. Ce taux est calculé en divisant le nombre de relations observées par le nombre maximum de relations possible. Un taux de liaison de 100% indique que l'un des deux termes de la relation est toujours présenté avec l'autre. Un taux de liaison proche de zéro indique que les deux termes ne sont pratiquement jamais présentés ensemble.

c) Résumé

Le logiciel Tropes est le seul représentant de sa catégorie. Il offre différents outils afin de procéder à une analyse cognitivo-discursive du corpus.
De part les fonctionnalités présentes, il peut toutefois être également classé dans d'autres catégories. En effet, il peut être mis dans les logiciels d'analyse socio-sémantique étant donné qu'il se base sur des dictionnaires et scénarios construits sur base d'une conception préalable des concepteurs. Tropes effectue également une analyse des cooccurrences ce qui permet de le classer parmi les logiciels d'analyse par réseau de mots associés. Enfin, dans la mesure où Tropes effectue le décompte de certaines unités lexicales regroupées par catégorie et par classes d'équivalents, il peut également être classé parmi les logiciels de lexicométrie.

L'une des particularités de Tropes est de procéder à une classification des unités lexicales selon différents critères. Ainsi, des termes peuvent être regroupés dans des classes d'équivalents s'ils sont sémantiquement proches. De même, chaque terme se voit affecté d'une catégorie (verbe, substantif, conjoncteur, etc.).

Une autre particularité de ce logiciel est de procéder directement à toutes les analyses dès l'ouverture d'un texte à analyser.

Tropes offre de nombreuses fonctionnalités / outils :

- La fonction « style » renvoie différentes analyses. Elle donne le style général du texte, sa mise en scène verbale, les propositions remarquables et les épisodes détectés.
- La fonction « Univers de référence » procède à l'identification des thèmes abordés dans le corpus en se basant sur les substantifs.
- La fonction « références utilisées » affiche les substantifs les plus significatifs utilisés dans le corpus regroupés sémantiquement en classes d'équivalents.
- L'outil « scénario » permet à l'utilisateur de créer ses propres thématiques. Le logiciel est équipé de « scénarios » préalables, mais il est possible de les modifier ou d'en créer de nouveaux.
- L'outil « relations » met en évidence les cooccurrences présentes dans le texte.
- L'outil « catégories fréquentes » affichent les catégories de mots significativement les plus fréquentes, tandis que l'outil « toutes catégories » affiche toutes les catégories de mots.
- La fonction « épisode » permet d'étudier la chronologie du discours et notamment de repérer les concentrations de certains termes dans certaines parties du texte.

3. Comparaison et évaluation des logiciels

Le dernier point de cette partie sera consacré à une comparaison des cinq logiciels présentés ci-dessus, ainsi qu'à leur évaluation.

La comparaison des logiciels portera principalement sur les points et fonctionnalités communs. En effet, lors de leur présentation, nous avons pu nous rendre compte de leur diversité. Certains outils ne sont donc pas comparables.

En ce qui concerne l'évaluation, différents critères ont été retenus.

Le premier critère sera la prise en main du logiciel. Est-il facile à aborder ? Faut-il plusieurs séances de manipulation pour arriver à s'en servir ou peut-on directement tirer profit des outils qu'il offre ? Nous rappelons que les logiciels choisis disposaient tous d'un manuel, critère qui a guidé le choix, et que cette caractéristique influencera nécessairement l'évaluation.

Le deuxième critère portera sur les prérequis nécessaires à l'utilisation du logiciel. Demande-t-il des connaissances préalables particulières ? Il apparaît évident qu'un minimum de connaissances en analyse de texte est toujours nécessaire.

Le troisième critère évaluera la gestion des requêtes. En dehors de la prise en main générale du logiciel, les outils d'analyse proposés par le logiciel sont-ils faciles à utiliser ? Nécessitent-ils de nombreux paramétrages ? Peuvent-ils seulement être personnalisés par l'utilisateur ?

Le quatrième critère portera sur la lisibilité des résultats. Les résultats affichés sont-ils faciles à comprendre ? Sont-ils parlants ?

Le cinquième critère évaluera la qualité des résultats. Correspondent-ils aux attentes des utilisateurs ? Ou réservent-ils des surprises ?

Enfin, le dernier critère sera la robustesse théorique. Le logiciel implémente-t-il la catégorie d'analyse de texte à laquelle il est censé se rattacher ?

Mais commençons d'abord par la comparaison des logiciels.

3.1. Comparaison

Dans ce point, nous allons procéder à une comparaison des outils proposés par les cinq logiciels présentés ci-dessus, lorsque cela est possible. En effet, certains logiciels offrent des fonctionnalités différentes des autres. Nous nous focaliserons principalement sur les fonctions communes.

Toutefois, avant d'aborder ces points communs, nous allons les comparer sur quelques critères plus généraux qui ont émergé de leur manipulation.

On peut s'interroger sur le degré de liberté laissé à l'utilisateur.

Une des grandes caractéristiques d'AntConc est de laisser au chercheur les décisions en matière d'analyse. AntConc ne réalise rien par lui-même. Il présente toujours des résultats bruts qui doivent être appréciés et triés pat l'utilisateur. Il appartient à l'utilisateur de décider des outils qu'il veut exploiter, et des paramètres de ces utilisations. Unitex laisse aussi une grande part de liberté. Il procède à l'indexation automatique du texte, mais les autres outils peuvent être utilisés au choix du chercheur.

À l'autre extrême, Tropes procède directement à toute une série d'analyses dès l'ouverture d'un fichier (détection des univers de référence, des relations, des épisodes, des rafales, etc.). Ce n'est que par la suite que l'utilisateur peut personnaliser certaines des fonctionnalités et notamment créer ses propres scénarios.

Entre les deux, les autres logiciels mélangent analyse automatique et choix de l'utilisateur. Lexico procède automatiquement à l'indexation du texte, mais l'utilisation des autres fonctionnalités doit être décidée par le chercheur qui peut, de plus, en paramétrer certains aspects. Sémato propose de manière automatique des thèmes, mais il appartient au chercheur de décider ou non d'utiliser cet outil et les autres fonctionnalités sont à sa libre appréciation.

Il convient également de faire la différence entre les logiciels apportant simplement une aide pratique au chercheur dans sa démarche d'analyse et ceux prétendant déjà procéder à une partie de l'analyse. Ce critère permet selon nous de classer les logiciels sur un continuum allant d'un extrême à l'autre.

AntConc nous a paru appartenir à la catégorie des logiciels se « contentant » d'apporter une aide pratique pour l'analyse. Ainsi, l'outil de concordance permettant d'afficher tous les contextes d'un terme a une visée purement pratique et visuelle. De même, l'indexation des termes, même si elle pose certaines questions[165], permet d'avoir une visualisation différente du texte. Unitex a également cette visée pratique. Même si l'indexation se fait sur base de dictionnaires et autres outils proposant déjà une forme d'analyse, les autres fonctionnalités permettent d'afficher certains éléments afin d'en avoir une vue différente.

A l'autre extrême, Tropes propose déjà des analyses à l'utilisateur : regroupement des termes en classes d'équivalents et univers, style du texte, propositions remarquables, etc. Il va donc plus loin qu'une simple aide pratique. Son objectif est d'offrir directement certains résultats au chercheur.

A nouveau, entre les deux, les autres logiciels apportent à la fois des outils pratiques et des résultats d'analyse. Lexico offre un concordanceur et une indexation du texte (aide pratique), mais il procède également à des analyses statistiques dont les résultats doivent être appréciés par l'utilisateur. De même, Sémato, par ses pages d'arrimage, permet d'organiser l'association de thèmes à un corpus. En même temps, il prétend proposer ses propres thèmes sur base des présupposés théoriques de ses concepteurs.

Ces premiers éléments posés, nous allons maintenant passer à la comparaison des fonctionnalités communes.

a) Indexation

L'indexation d'un texte procède au listage de toutes les unités lexicales qu'il contient.
Cette fonctionnalité est présente chez AntConc, Lexico, Tropes et Unitex.

AntConc procède à l'indexation sur base du repérage de caractères précis que l'utilisateur peut personnaliser et des blancs. Par défaut, une unité lexicale est composée des lettres de l'alphabet. Mais il est possible d'y ajouter notamment les caractères de ponctuation et les nombres. Il affiche également la fréquence des unités lexicales.

Lexico procède à l'indexation sur base de la reconnaissance de caractères délimiteurs personnalisables par l'utilisateur, les autres caractères composant les unités lexicales. Il effectue cette opération dès l'ouverture d'un texte. Il affiche également les fréquences.

[165] Ambiguïtés, lemmatisation, etc.

Lexico ne procède donc pas à une lemmatisation du texte, c'est-à-dire à un regroupement des formes fléchies avec leur équivalent canonique. Il est toutefois possible pour l'utilisateur de le faire lui-même en regroupant les termes qu'il souhaite voir considérés ensemble.

Cela signifie donc aussi que Lexico ne repérera pas les termes inconnus de la langue.

Unitex procède également à une indexation du corpus sur base des dictionnaires incorporés au logiciel, mais également par le repérage de séparateurs. Les fréquences des unités lexicales sont affichées, ainsi que leur répartition en mots simples, mots composés et mots inconnus.

En ce qui concerne Tropes, l'indexation se fait de manière particulière. Les mots sont regroupés en catégories (substantifs / références, verbes, connecteurs, modalisations, pronoms) et parfois en sous catégories (exemple : modalisations de temps, de lieu, etc.).
En ce qui concerne les références, seules les plus significatives sont retenues. Pour chaque sous-catégorie, la fréquence est indiquée. Cette indexation se base sur un dictionnaire présent dans le logiciel et Tropes procède à une lemmatisation des termes.

Les approches avec ou sans dictionnaire présentent chacune des avantages et des inconvénients, les avantages de l'une se présentant souvent comme les inconvénients de l'autre. La lemmatisation ne peut avoir lieu qu'en présence d'un dictionnaire. Mais elle exige une analyse assez fine afin de déterminer correctement la forme canonique. Des erreurs sont toujours possibles.
Par exemple, si l'on prend le texte « je livre », la présence du pronom « je » permettra de rattacher « livre » au verbe « livrer ». Si l'on a « le livre », on se tournera plutôt vers le substantif « livre ». Mais comment procéder vis-à-vis de « je le livre » ?
En l'absence de dictionnaire, les termes ne pourront être regroupés que sur leur forme graphique. Les homonymes seront donc regroupés ensemble.
Comme on le voit, le choix de l'une ou l'autre implémentation ne permet pas de résoudre tous les problèmes et d'arriver à une indexation fiable à cent pour-cent.

b) Balisage du texte

Le balisage d'un texte permet de découper celui-ci en différentes parties afin de procéder à des comparaisons entre celles-ci.

Cette fonction est présente chez Lexico et est un préalable obligatoire à l'utilisation des fonctionnalités statistiques avancées. Il appartient à l'utilisateur de définir ses propres balises et donc son découpage.

Tropes permet également d'introduire des délimiteurs dans le texte, mais uniquement dans sa version payante.

Le logiciel Sémato procède à un découpage automatique du texte sur base des retours chariot. Il appartient donc à l'utilisateur d'introduire ceux-ci aux endroits adéquats.

c) Concordance

La concordance qui consiste à afficher le contexte d'un terme ou d'une expression est présente chez AntConc, Lexico, Tropes et Unitex.

Elle constitue une forme de recherche.

AntConc et Lexico permettent de personnaliser les résultats (tri, taille du contexte).

Unitex permet également l'affichage de concordances de termes ou expressions rationnelles.

Chez Tropes, cette fonctionnalité est accessible depuis les résultats des différentes fonctionnalités. Par exemple, il est possible de cliquer sur un terme d'un univers de référence pour que s'affichent toutes les phrases qui le contiennent.

d) Outils quantitatifs

Les outils quantitatifs correspondent à la lexicométrie et aux analyses statistiques.
Les différents logiciels en offrent.

Via son outil d'indexation, AntConc donne la fréquence absolue des unités lexicales présentes dans le corpus. Il procède également au décompte des cooccurrences.

Lexico offre de nombreux résultats quantitatifs. Il affiche le décompte des unités lexicales lors de l'indexation (fréquence absolue), des segments répétés. Il affiche également des statistiques générales sur le texte et ses parties : nombre d'occurrences, nombre de formes, nombre d'hapax, fréquence maximale, etc. Il permet d'analyser la répartition quantitative (fréquence absolue, fréquence relative, spécificités) de termes entre les différentes parties d'un texte en cas de balisage de celui-ci. Il procède également à une analyse factorielle des correspondances.

Unitex présente la fréquence des unités lexicales.

Les outils de recherche de Sémato présente des résultats chiffrés : fréquences d'unités textuelles, fréquences de phrases, réseau de similitudes, etc.

Tropes produit également les fréquences des résultats. Il permet également de détecter des rafales, fonction comparable aux spécificités de Lexico, c'est-à-dire détection de termes sur- ou sous-représentés dans certaines parties du texte.

e) Cooccurrences

La recherche de cooccurrences, c'est-à-dire de termes fréquemment associés est présente chez AntConc, Lexico, Sémato, Tropes.

AntConc offre deux outils de recherche des cooccurrences. Le premier recherche des « clusters », c'est-à-dire les termes immédiatement voisins (avant ou après au choix) d'un terme précisé par l'utilisateur. Le deuxième permet de rechercher les termes apparaissant dans le contexte d'un terme défini par l'utilisateur, ce dernier devant préciser la taille dudit contexte. Les résultats sont personnalisables.

L'outil Lexico offre un outil qui peut s'apparenter à la recherche de cooccurrences.
L'outil « segment répété » identifie dans le texte des suites de formes dont la fréquence est supérieure à 2. Il peut, par exemple, identifier des expressions comme « président de la République », « moulin à café », c'est-à-dire des suites de termes, des termes cooccurrents, qu'il est intéressant de considérer ensemble.

Sémato permet également de rechercher des cooccurrences entre thèmes, et non entre termes, associés à un texte. Il en affiche la fréquence.

Tropes est équipé d'un outil visant à détecter les cooccurrences entre les classes d'équivalents détectées. Leur fréquence est affichée.

f) Recherche

Les outils de recherche recouvrent des fonctionnalités assez diverses. En soi, l'outil concordance vu ci-dessus est une forme de recherche dont l'objectif est d'afficher le contexte d'un terme. De même, l'outil cooccurrence permet de chercher des fréquences d'association de termes.

En plus de ces exemples, différents logiciels offrent des outils de recherche plus moins sophistiqués.

Sémato, via son outil « Requête en repérage », permet de faire différentes recherches dans le corpus. Il est possible de rechercher des termes précis ou expression, Sémato renvoyant la liste des unités textuelles contenant l'objet de la recherche. Il est également possible de rechercher les unités textuelles correspondant à des thèmes introduits manuellement ou automatiquement. Ces recherches peuvent être personnalisées selon différents critères.

Chaque texte transféré sur le serveur de Sémato peut être personnalisé par des catégories. Il est possible, via l'outil « Analyse », de faire une recherche croisée entre catégories de texte et thèmes définis. Quant à l'outil « réseaux de similitudes », il permet de repérer les unités textuelles se ressemblant fortement.

Unitex, quant à lui, permet des recherches assez poussées de motifs complexes dans un corpus, notamment par l'utilisation de graphes.

g) Thématisation

La thématisation vise les outils qui ont pour objectif d'identifier automatiquement les thèmes qui seraient présents dans un corpus.

Sémato, via son outil « Thème », propose automatiquement des thèmes par rapport à un corpus. Il appartient ensuite à l'utilisateur de retenir ou non ceux-ci et / ou de les personnaliser. Ceux-ci sont subdivisés en trois catégories : objets (ce dont on parle), actions (ce que l'on fait) et qualités (qualifications).

Pour produire ces thèmes, Sémato attribue à chaque mot un champ sémantique, c'est-à-dire « *une liste d'autres mots trouvés dans le corpus et apparentés sémantiquement* »[166].
L'ensemble des champs sémantiques du corpus forme un réseau.
« *Des procédures de **Sémato** permettent de fusionner des zones du réseau, répondant à certaines caractéristiques mathématiques, afin de proposer de nouvelles catégories. En **Sémato**, les propositions de catégorisation sont émergentes de ce réseau et non parachutées depuis un dictionnaire.* » [167]

[166] http://fable.ato.uqam.ca/guidexpert-ato/gea-top.asp
[167] http://fable.ato.uqam.ca/guidexpert-ato/gea-top.asp

Tropes, via l'affichage d'Univers de références, vise à identifier les thèmes du corpus qui lui est soumis. Pour rappel, Tropes regroupe les substantifs en classes d'équivalents sur base d'une proximité sémantique. Ensuite, ces classes sont elles-mêmes regroupées selon deux niveaux (univers de référence 1 et 2). Chaque univers est affecté de sa fréquence. Pour effectuer cette analyse, Tropes se base sur un dictionnaire d'équivalents sémantiques partiel (tous les mots du français n'y figurent pas).

Une autre manière d'identifier des thèmes est de créer ceux-ci via l'outil scénario et de l'appliquer au corpus afin de voir s'ils sont présents. Un scénario permet d'établir une arborescence de termes, et d'effectuer des regroupements sémantiques sur plusieurs niveaux. Tropes est équipé de scénarios, plus ou moins détaillés, qui peuvent être appliqués au corpus.

Tableau récapitulatif

Logiciels/ fonctionnalités	AntConc	Lexico	Unitex	Sémato	Tropes
Indexation	+	+	+		+
Balisage		+		+	+
Concordance	+	+	+		+
Outils quantitatifs	+	+	+	+	+
Cooccurrences	+	+		+	+
Recherche	+	+	+	+	+
Thématisation				+	+

3.2. Évaluation

Dans ce deuxième point, nous allons procéder à une évaluation des différents logiciels analysés.

Comme cela a déjà été indiqué dans l'introduction de ce point, nous avons retenu différents critères :

- Prise en main : le logiciel est-il facile à aborder ?
- Prérequis : le logiciel demande-t-il des connaissances particulières ?
- Gestion des requêtes : les outils d'analyse proposés par le logiciel sont-ils faciles à utiliser ?
- Lisibilité des résultats : les résultats affichés sont-ils faciles à comprendre ?
- Qualité des résultats : les résultats correspondent-ils aux attentes ?
- Robustesse théorique : le logiciel répond-il à la catégorie d'analyse à laquelle il est censé se rattacher ?

a) AntConc

La prise en main de AntConc est assez facile. Le nombre de fonctionnalités est limité et le manuel d'explication, en anglais mais assez court, explique rapidement les attentes par rapport à chacune. Chaque fonctionnalité est indiquée par un onglet facilement repérable. Les options des fonctions sont aussi bien indiquées au-dessous de l'écran. Nous lui mettrons donc 9/10 pour ce point.

Les prérequis exigés pour pouvoir utiliser AntConc sont relativement faciles à acquérir. Il convient d'avoir quelques notions de base en analyse de texte (lexicométrie, analyse des

cooccurrences), ainsi que certaines connaissances informatiques (expressions régulières). Nous lui mettrons 8/10.

En ce qui concerne la gestion des requêtes, celles-ci sont assez facilement paramétrables. Le bas de l'écran affiche tous les éléments qui peuvent être personnalisés par l'utilisateur et leur prise en main est rapide. Rappelons qu'il est possible de « charger » plusieurs textes dans le logiciel afin d'effectuer des requêtes parallèles. Nous lui mettrons 9/10.

Les résultats sont généralement affichés sous forme de listes avec fréquence et rang. Leur lisibilité est très facile. Seul l'outil « Concordance plot » qui affiche les résultats sous forme d'un code barre nous a semblé moins significatif. Il est en effet difficile de se représenter la place exacte d'un terme seulement en le regardant. Nous mettrons 8/10.

Tous les résultats affichés sont bruts, que ce soit en terme de concordance, d'indexation, de cooccurrences, etc. Il n'y a aucune sélection, aucune interprétation quant à ceux-ci. Leur qualité correspond donc aux attentes. Nous lui mettrons 8/10.

Enfin, au niveau de la robustesse théorique, AntConc permet de faire de la lexicométrie et de l'analyse des cooccurrences. Se présentant plutôt comme un outil d'aide que d'analyse, il laisse les phases d'inférence et d'interprétation au chercheur. Il ne contredit donc pas la théorie. Nous lui mettrons 9/10.

b) Lexico

La prise en main de Lexico est un peu plus complexe, mais reste abordable relativement aisément. Le manuel est généralement d'une bonne aide. Pour rappel, l'utilisation des outils statistiques demande un balisage du texte. Il convient donc de préparer celui-ci auparavant. On peut aussi regretter que deux fonctionnalités ne soient pas expliquées dans le manuel. Nous lui mettrons 7/10.

Les prérequis exigés pour pouvoir utiliser Lexico sont plus importants : connaissance de base en analyse de texte (lexicométrie, cooccurrence), mais surtout connaissance de certains outils statistiques. Nous lui mettrons 7/10.

La gestion de requêtes est variable. La recherche de concordances ou de segments répétés, ou la création de groupes de formes sont faciles à utiliser. L'utilisation des outils statistiques est plus complexe, les paramètres à configurer demandant une connaissance pointue de ces outils. Nous lui mettrons 6/10.

Les mêmes considérations qu'au point précédent peuvent s'appliquer ici aussi en ce qui concerne la lisibilité des résultats. Les affichages par liste sont faciles à aborder. La lecture des graphiques statistiques demande des connaissances avancées. Nous mettrons 6/10.

La qualité des résultats demande pour certains outils des connaissances pointues. Certains graphiques produits sont difficiles à interpréter ou à estimer sans référence préalable. De plus, l'indexation se basant sur les formes graphiques, les termes ambigus ne seront pas dissociés. Nous mettrons 7/10.

Au niveau de la robustesse théorique, Lexico est un représentant de la lexicométrie et des analyses statistiques textuelles. Il implémente certains outils statistiques, mais pas d'autres. Déjà à sa troisième version, il fera sans doute encore l'objet d'amélioration et d'ajouts de la part de ses concepteurs dans le futur. Nous lui mettons 8/10.

c) Unitex

La prise en main d'Unitex a été plus difficile. Le manuel est relativement long. Les fonctions simples telles l'ouverture d'un fichier pour indexation ou la construction du graphe du texte sont faciles à aborder. Par contre, l'utilisation des différentes fonctionnalités de recherche (par graphe) est beaucoup plus complexe. Nous lui mettrons 5/10.

Au niveau des prérequis, Unitex demande des connaissances dans différents domaines : expressions régulières, grammaire, graphes, etc. Nous lui mettrons 6/10.

La gestion des requêtes est variable. Certains outils sont assez difficiles à utiliser (recherche par graphe, par expressions régulières) et présentes des « bugs » impossibles à corriger. Nous lui mettrons 4/10.

Les résultats affichés sont par contre assez lisibles. Ce sont généralement des listes ou des graphes. Nous lui mettrons 8/10.

En ce qui concerne le contenu des résultats et leur qualité, certaines fonctions (indexation, graphe du texte) sont basées sur des dictionnaires. Ceux-ci ont pour objectif de résoudre les problèmes pouvant se présenter (ambiguïtés, lemmatisation). Les résultats sont donc dépendant de leur contenu. Toutefois, on peut noter que le graphe du texte fait apparaître toutes les solutions possibles à l'utilisateur. Nous lui mettrons 7/10.

Enfin, en ce qui concerne la robustesse théorique, Unitex permet de faire du simple décompte de termes, donc de la lexicométrie de base. Les autres fonctions sont surtout axées sur la recherche. Les résultats sont relativement bruts et il appartient au chercheur de tirer ses propres conclusions. Nous mettrons 8/10.

d) Sémato

La prise en main de Sémato présente tant des aspects aisés que complexes. L'utilisateur est guidé par un manuel disponible en ligne. Si le transfert de fichiers sur le serveur du logiciel est aisé, ainsi que l'utilisation de la plupart des outils, certains sont moins faciles à appréhender. De plus, Sémato utilise le terme « texte » pour paragraphe, ce qui est déroutant. Nous lui mettrons 7/10.

Au niveau des prérequis, il est nécessaire d'avoir des connaissances générales en matière d'analyse de texte (cooccurrences, thèmes). La recherche d'une Configuration focus demande toutefois des connaissances plus pointues. Nous lui mettrons 7/10.

En ce qui concerne les requêtes, elles demandent parfois des nombreux paramétrages, parfois trop. Ce qui est une force, peut parfois devenir un obstacle. Nous mettrons 7/10.

Les résultats sont généralement très lisibles. Ils se présentent sous formes de listes de textes répondant à une requête spécifique. Toutefois, l'outil Configuration focus est beaucoup plus complexe et les résultats qu'il produit ne sont pas évident à lire.
Il est également possible d'avoir des résultats sous forme graphique, facilement abordables. Nous mettrons 7/10.

La qualité des résultats est variable. Toutes les requêtes simples (recherche de termes par exemple) correspondent aux attentes. L'outil de thématisation automatique peut réserver des

surprises, celui-ci étant basé sur un présupposé théorique des concepteurs. Nous mettrons 6/10.

Enfin, au niveau de la robustesse, Sémato correspond à un outil d'analyse textuelle socio-sémantique, c'est-à-dire qu'il se base sur des concepts préalables, principalement en matière de thématisation automatique, dont il convient d'apprécier la pertinence. Malheureusement, tous les renseignements nécessaires à la correcte appréciation de ceux-ci n'ont pu être trouvés. Pour cette raison, nous lui mettrons 4/10.

e) Tropes

La première prise en main de Tropes est assez aisée puisque le logiciel affiche un court texte indiquant comment procéder. Immédiatement, toutes les fonctionnalités de base sont exécutées sans que l'utilisateur n'ait à intervenir. Tropes est de plus équipé d'un manuel très complet. Enfin, des onglets d'explication sont présents pour chaque résultat. La prise en main de l'outil « Scénario » est toutefois plus complexe. Nous lui mettrons 8/10.

Les prérequis nécessaires pour pouvoir utiliser Tropes sont plus importants. En seul représentant de l'analyse cognitivo-discursive, il demande une connaissance de cette théorie qui n'est pas facile à aborder. Il en est de même pour les notions de styles, graphes, rafales, etc. Nous mettrons 6/10.

Au niveau de la gestion des requêtes, la plupart des outils sont exécutés automatiquement. L'utilisateur a peu de choix. Certains paramétrages peuvent être faits via le menu général. L'outil qui demande la plus grande manipulation par l'utilisateur est « Scénario ». Celui-ci demande la construction d'une arborescence terminologique qui n'est pas toujours évidente. Nous mettrons 6/10.

La lisibilité des résultats est assez variable. L'aide du manuel et des onglets d'explication est précieuse pour interpréter les différents graphes et certains résultats (styles, épisodes, rafales). Nous mettrons 7/10.

La qualité des résultats laisse parfois à désirer. Par exemple, les scénarios pré-équipant Tropes se basent sur une certaine conception des auteurs quant à la classification et aux liens unissant le vocabulaire, ce qui peut parfois amener des surprises. Ainsi, lors de l'utilisation de Tropes sur un autre texte, le terme « branche » dans l'expression « branche du droit » avait été placé en botanique. Face à des termes du langage courant utilisés avec une acception différente dans certains domaines, les scénarios de base peuvent amener à des incohérences. Il en est de même avec les univers de référence qui reposent également sur une classification et une sélection préalable des concepteurs. Nous mettrons 5/10.

Enfin, en ce qui concerne la robustesse théorique, il convient en premier lieu de noter que l'analyse cognitivo-discursive nous semble assez subjective. Reposant sur la détection des substantifs clés d'un corpus, les critères permettant de les identifier ne nous ont pas parus clairs. De même, la classification des mots en catégories et sous-catégories peut parfois être sujette à discussion. Il convient également de noter que si l'analyse cognitivo-discursive est une méthode qualitative, Tropes n'a pas pu faire l'impasse sur des calculs statistiques pour certaines fonctionnalités.

Il est difficile de juger ce logiciel sur ce critère étant donné que Tropes a été conçu par ou en collaboration avec l'auteur de cette théorie d'analyse textuelle. Les deux sont donc intimement liés. Nous mettrons 4/10.

f) Tableau récapitulatif

Le tableau suivant reprend les notes attribuées aux logiciels et en fait la somme.

	Prise en main	Prérequis	Gestion des requêtes	Lisibilité des résultats	Qualité des résultats	Robustesse théorique	total
AntConc	9	8	9	8	8	9	51
Lexico	7	7	6	6	7	8	41
Unitex	5	6	4	8	7	8	38
Sémato	7	7	7	7	6	4	38
Tropes	8	6	6	7	5	4	36

Leur interprétation doit être abordée avec prudence. Ils résultent de la seule utilisation de l'auteur de ce mémoire.

La conclusion générale qui peut être tirée de ces résultats est que les outils les plus simples, qui apportent surtout une aide pratique, sont sans doute les plus fiables. Il laisse toute l'interprétation au chercheur. Les outils suppléant le chercheur dans certaines analyses reposent sur des présupposés théoriques qui ne seront pas nécessairement ceux de l'objet de la recherche. Dans ce cas, il appartient au chercheur de s'interroger sur l'adéquation des concepts qui se cachent derrière le logiciel afin de déterminer s'ils sont pertinents dans le cadre de sa recherche.

Chapitre 4 : Exigences des chercheurs et choix de quelques logiciels

Les recherches effectuées dans le cadre de ce mémoire avaient déjà laissé sous-entendre que le domaine de l'analyse de texte était l'apanage des sciences humaines, sociales, économiques et littéraires. En effet, l'utilisation de ces logiciels semble être principalement le fait de personnes travaillant dans des domaines tels que la sociologie, la psychologie, la littérature, etc., mais rarement (voire jamais) de domaines plus techniques ou scientifiques (ingénierie, chimie, médecine, etc.) et donc, de l'informatique.

Les interviews récoltées confirment en partie ce constat. En effet, si la démarche d'analyse de texte n'est pas absente du monde informatique, elle n'est toutefois pas prééminente et est de plus très ciblée.

Quatre personnes ont été interviewées. Trois appartiennent au « monde informatique ». La quatrième procède à des recherches de type sociologique.

De cette différence de domaine, est ressortie une différence quant aux besoins en matière d'analyse de texte.

L'objectif de cette partie est de faire une analyse approfondie des interviews réalisées dans le cadre de ce mémoire afin d'identifier les attentes et exigences des chercheurs en matière d'analyse de texte.

Les différents cas d'analyse de texte cités ont un objectif précis. L'atteinte de ces objectifs suit une méthodologie précise et recourt à des fonctionnalités et outils dont certains peuvent être du ressort des logiciels d'analyse de texte.
Il s'agira donc de présenter pour chaque objectif, les outils qui pourraient être utiles.

Nous allons commencer par analyser chaque interview en elle-même afin de cerner les éléments clés.

Ensuite, nous procèderons à une synthèse des éléments recueillis afin de pouvoir détecter les similitudes et les divergences.

1. Analyse de l'interview de Michaël Petit

Lors de son interview, Michaël Petit a passé en revue les différents cas où il se trouve confronté à une analyse de texte. Comme nous le verrons dans les lignes qui suivent, celle-ci est entendue au sens large. Il avait été décidé au départ de ne pas trop orienter l'interviewé afin de recueillir le maximum d'enseignements.

- Objectif : contrôler le plagiat

La première demande concerne des « *logiciels permettant de contrôler le plagiat dans les travaux d'étudiants* ».

Le plagiat consiste en une « reproduction non avouée d'une oeuvre originale ou d'une partie de cette dernière ».[168]

[168] http://atilf.atilf.fr/dendien/scripts/tlfiv5/visusel.exe?35;s=2974718055;b=7;r=1;nat=;i=2;;

Si la définition semble claire, elle peut cependant soulever certaines questions. À partir de quand y a-t-il plagiat ? Il est évident que la reproduction de phrases ou paragraphes entiers en fait partie. Bien que l'on puisse hésiter pour des phrases tellement simples qu'il paraît impossible de les formuler autrement. L'on peut aussi avoir des doutes lorsque l'on reprend des expressions. Il convient de déterminer si celles-ci sont du ressort exclusif de l'auteur cité ou font partie du vocabulaire du thème.

Un logiciel permettant de détecter le plagiat devrait être capable de comparer des textes entre eux en signalant les phrases, portions de phrases, paragraphes, voire sections entières qui seraient communs. Nous appellerons cette fonction « **identités entre textes** ».

Il appartiendrait alors au professeur de vérifier si les parties reprises ont été mentionnées avec leur origine. Mieux encore, le logiciel idéal devrait pouvoir repérer si ces parties communes ont été marquées comme citations (présence de guillemets et de notes de bas de page par exemple).

Un tel logiciel devrait de plus pouvoir disposer d'une base de données complète concernant le sujet traité dans le travail d'étudiant. Celle-ci pourrait être fournie avec le logiciel et / ou alimentée par l'utilisateur.

- Objectif : rechercher des sites et documents sur le web

Une autre demande a trait à des « *logiciels permettant de faire des recherches sur un sujet particulier* » sur le web.
On se situe ici dans une « *démarche quantitative* ».

Via l'introduction d'un ou plusieurs mots-clés, le logiciel devrait pouvoir afficher une liste de sites ou documents pertinents (fonction « **recherche et sélection sur le web** »).

Cette fonction s'apparente aux moteurs de recherche Google et autres. L'inconvénient de ceux-ci étant de ne pas toujours sortir les sites les plus adaptés. Il conviendrait donc d'avoir un logiciel faisant une analyse plus fine des sites et des documents afin d'identifier les sites pertinents.
Ce dernier devrait également permettre de chercher « *par des synonymes ou des mots connexes du thème que l'on veut explorer* ». Le logiciel devrait être équipé d'un dictionnaire ou thésaurus de synonymes, éventuellement organisé de manière thématique (outil « **dictionnaire** »).

- Objectif : organiser des documents

Une autre attente concerne des « *logiciels d'indexation des documents du disque dur* ».

L'objectif serait de « *retrouver un document attaché à un mail ou rangé à un endroit auquel on ne pense plus* ». Toutefois, « *L'organisation ne peut rester que manuelle* ».

Nous nous situons cette fois-ci dans une « *perspective quantitative, mais éventuellement étendue à une base qualitative / syntaxique* ».

Au niveau de la méthodologie qui pourrait être suivie, il s'agirait dans un premier temps d'identifier les mots clés du texte, reflets des thèmes présents, afin de pouvoir classer celui-ci.

Afin de pouvoir repérer le vocabulaire clé d'un texte, un logiciel doit d'abord pouvoir indexer celui-ci (fonction « **indexation** »).

L'indexation consiste dans le « fait d'établir un relevé complet du vocabulaire, c'est-à-dire un inventaire de toutes les formes ou des unités lexicales qui figurent dans un énoncé ou un ensemble d'énoncés déterminé(s) ».[169]

En tant que tel, l'inventaire des unités lexicales pose certaines questions.
Il convient déjà de définir à quoi ce concept correspond. Doit-on retenir chaque forme différente (un singulier est différent d'un pluriel, un verbe conjugué d'un infinitif) ou procéder à une **lemmatisation** (regroupement des formes singulier et pluriel, conjugué et infinitif, etc.) ?
L'identification des termes composés ou d'expressions pose également certains problèmes. La présence d'un tiret « - » ne suffit pas toujours (contre-exemple : *doit-on* …).
La gestion des homonymes ou ambiguïtés (**gestion des ambiguïtés**) n'est pas non plus évidente. Elle suppose une analyse sémantique du terme sur base de son contexte.

Il est également demandé que ces logiciels puissent gérer le « *problème des mots connexes*[170] », fonction « **connexité** », c'est-à-dire de regrouper ensemble des unités lexicales fortement liées.

L'indexation faite, il convient de sélectionner les termes qualifiant le texte.
Une solution peut être de présenter pour chaque texte un tableau des unités lexicales, éventuellement accompagnées de leur fréquence, et de permettre à l'utilisateur de sélectionner les unités lexicales à garder. Par exemple, les articles définis « la, le, les, l' » ne sont pas nécessairement utiles.
Une autre solution est de laisser le logiciel choisir lui-même les unités lexicales pertinentes (fonction « **thématisation** »).

- Objectif : créer des modèles

Une autre forme d'analyse de texte concerne la « *création de modèles* » sur base d'interviews ou d'observations d'acteurs sur le terrain.

Ces analyses se font généralement manuellement. Il s'agit par exemple de « *créer des diagrammes de séquence et de là, des diagrammes d'activité* ».
« On extrait des connaissances pour les traduire en modèle UML. »
« *On observe des scénarios : des séquences de choses. Et on essaie de généraliser à partir de plusieurs.*
Ce n'est pas thématique, c'est plutôt séquentiel, une série d'étapes. Il y a les acteurs, les activités observées et les données manipulées. »
« *A partir de documents en langage naturel, il s'agit d'extraire des éléments pour en faire un modèle structuré.* »

[169] http://atilf.atilf.fr/dendien/scripts/tlfiv5/visusel.exe?12;s=1754764905;r=1;nat=;sol=1;
[170] « Qui est en relation étroite (avec) »,
http://atilf.atilf.fr/dendien/scripts/tlfiv5/advanced.exe?50;s=1754764905;

Un logiciel de création de modèles devrait présenter des fonctionnalités très spécifiques et avancées. Il devrait pouvoir, à partir du texte, identifier les éléments du modèle, éventuellement sur base de mots clés.

- Objectif : appliquer un modèle

Dans le même ordre d'idée, il est demandé des « *logiciels qui pourraient fournir des schémas entité-association* » ou « *décrire les fonctions attendues d'un système en analysant les exigences mentionnées dans des interviews* ».

A partir d'un langage de modélisation existant, il s'agirait pour le logiciel d'analyser des textes et de procéder à une modélisation de ceux-ci en fonction, par exemple, du vocabulaire utilisé ou de certaines tournures de phrases.

Le logiciel devrait être équipé d'un algorithme de repérage des éléments du modèle. Il devrait être possible d'alimenter cet algorithme au fur et à mesure des analyses afin de couvrir de plus en plus de cas.

Par exemple, dans le cadre de schémas entités-associations, le repérage des attributs des entités pourrait se faire via l'utilisation des verbes « avoir », « possède », etc.

Une aide pourrait déjà être apportée par des outils de **recherche** permettant de retrouver rapidement des passages de textes au sein d'un corpus volumineux.

Ce genre de fonction demande une analyse approfondie du texte et la résolution de nombreux problèmes (ambiguïtés, etc.).

- Objectif : contrôler le respect d'un template

Une autre utilisation de logiciels d'analyse de texte pourrait concerner les cahiers des charges. Il s'agit de textes très structurés qui utilisent des templates pour leur rédaction.
« *En ce qui concerne les logiciels d'analyse, on pourrait vérifier que le contenu d'une section est adapté à ce qui est demandé via le vocabulaire utilisé.* ».

Il s'agirait pour le logiciel d'analyser les différentes parties du texte en vérifiant la correspondance du lexique avec le vocabulaire attendu en fonction du template utilisé.

Le logiciel devrait avoir en mémoire les différentes structures de template possibles. A chaque structure serait associé un vocabulaire type, réparti en fonction des sections divisant la structure. Ce vocabulaire devrait être mis dans des dictionnaires spécialisés.

Ce vocabulaire devrait être assez large pour pouvoir couvrir un maximum de cas et envisager les synonymies possibles.
Il devrait aussi pouvoir être enrichi par l'utilisateur au fil des utilisations.

Au niveau méthodologique, le logiciel commencerait par un découpage du texte correspondant aux différentes sections du template. Ces sections pourraient être identifiées sur base de mots clés (chapitre, paragraphes, etc.) ou de balises introduites préalablement.

En même temps, il procèderait à une indexation du texte par parties en mettant en évidences les termes appartenant aux dictionnaires spécialisés, ainsi que leur fréquence.

Il appartiendrait alors au chercheur d'interpréter ces résultats.

- Objectif : structurer une ontologie

Une autre utilisation de logiciels d'analyse de texte pourrait avoir lieu dans le domaine de l'ontologie. A ce sujet, différentes attentes peuvent être identifiées.

Une première attente concerne la structuration d'une ontologie. « *L'analyse doit permettre de détecter des concepts que l'on a dans le modèle. Cela facilite la transition vers un langage plus formel*. » Le logiciel doit offrir un « *outil d'analyse dirigé par un vocabulaire déjà existant* ». Il doit permettre de « *structurer les connaissances que l'on a déjà grâce à un outil* »

Une ontologie est « *un ensemble de concepts hiérarchisés et ayant des relations entre eux* ».

Il s'agit pour le logiciel de partir d'un vocabulaire existant et d'aider le chercheur à structurer celui-ci en une ontologie. Il doit pouvoir présenter ce vocabulaire et les liens entre les termes.

Au niveau des fonctionnalités que doit présenter un tel logiciel, on peut penser à l'« **indexation en parallèle** » qui permettra de repérer les concepts préexistants dans les textes.
Il devrait également permettre d'« **afficher en parallèle les contextes d'un ou plusieurs termes** ».

- Objectif : créer une ontologie

Une autre fonction en rapport avec l'ontologie concerne la création d'une ontologie. Contrairement à la fonction précédente, on ne part pas d'un vocabulaire existant, mais l'on doit identifier celui-ci.

Des « *logiciels seraient aussi utiles pour créer une description d'un domaine de recherche, une ontologie reprenant tous les concepts d'un domaine et leurs liens* ».

« L'analyse de texte devrait permettre d'extraire tous les concepts importants du domaine sur base d'articles de personnes de l'équipe ou d'autres articles.
Le logiciel devrait pouvoir extraire les termes et leurs relations afin de faire l'ontologie du domaine. »

Traditionnellement, « *On peut procéder de deux manières pour faire cette ontologie.*
On peut réfléchir dessus et la définir manuellement.
Ou on peut analyser les textes existants sur le domaine. On prend l'ensemble des documents des spécialistes du domaine et on procède à une extraction des termes pertinents via un logiciel d'analyse. »
« *Il s'agit à la fois d'une démarche qualitative et d'analyse des relations entre les concepts.* »

La création d'ontologies se base sur des « *mesures classiques* ».

Un logiciel de création d'ontologies devrait offrir différentes fonctionnalités. Il devrait d'abord procéder à l'« **indexation en parallèle** » des textes du domaine (avec les problèmes que cela soulève, voir avant). Il devrait également présenter la fréquence de chaque unité ou d'autres indications statistiques (« **lexicométrie** »), celles-ci ne devant être considérées que comme des indices pour le chercheur.

Un aspect très important de ce domaine est le « *contraste entre vocabulaires* ». « *Un terme n'est pertinent pour un domaine que s'il est présent dans les articles du domaine et absent dans les articles d'autres domaines* » (fonction « **contraste** »).

Un tel logiciel devrait également inclure un « *outil permettant de retrouver les définitions des termes dans les articles ou ailleurs* » (fonction « **définition** »).

Ensuite, il devrait offrir un outil de « **présentation de la taxonomie** ». « *Ces termes seraient placés dans une taxonomie*[171] *(du général vers le particulier), par inclusion des termes particuliers dans un terme plus général.* »

Enfin, il devrait permettre la **recherche d'informations liées** à la taxonomie, par exemple des sites web. « *Un deuxième outil devrait permettre de découvrir des informations liées à l'ontologie.* » Cela renvoie à un des objectifs identifiés ci-avant.

- Objectif : enrichir une ontologie existante

Enfin, une dernière fonction en rapport avec l'ontologie a été citée : l'enrichissement d'une ontologie. « *De plus, il devrait être possible d'enrichir l'ontologie à partir des nouvelles découvertes.* »

Il s'agit d'enrichir l'ontologie existante sur base de nouveaux textes incorporés à la base de données existante. Ceux-ci sont analysés afin de déterminer si de nouveaux concepts peuvent être intégrés ou si de nouveaux liens entre termes peuvent être établis.

Dans ce cadre, différentes fonctions peuvent être présentes.
« *Le logiciel devrait pouvoir analyser la pertinence du document et l'intégrer dans l'ontologie.* » Il s'agit de vérifier la **pertinence du texte** par rapport au domaine étudié.
« *Il devrait pouvoir faire le matching entre l'ontologie et l'information textuelle brute.* »

Les autres fonctions correspondent à celle de la création d'une ontologie.

À nouveau, le problème de l'ambiguïté des termes est présent. Il peut d'ailleurs s'appliquer aux différents aspects de l'ontologie.
« *Mais tout ne pourrait pas être automatique car un mot-clef peut avoir plusieurs significations et il peut y avoir plusieurs mots pour un concept. Il faudrait des systèmes intelligents avec détection des synonymes et des relations entre concepts.* »

[171] Une taxonomie est une classification sous forme d'arbre sans cycle et linéaire, on descend dans l'arbre du plus général au plus spécial. Dans une ontologie, il y a des liens entre les concepts en plus.

2. Analyse des interviews de Patrick Heymans et Nicolas Mayer

L'interview de Patrick Heymans était plus orientée vers l'activité « recherche ».

Celle-ci a été complétée par celle de Nicolas Mayer.

- Objectif : uniformiser une ontologie existante

Une première attente vis-à-vis d'un logiciel d'analyse de texte serait l'uniformisation d'une ontologie.

L'interviewé s'est basé sur un exemple précis, mais il pourrait s'appliquer à n'importe quel domaine.

La problématique est expliquée comme suit.
« *L'état de l'art est très riche mais divergent.*
D'un côté, on a les standards d'analyse des risques venant des organisations de standardisation (surtout des industriels). Ces documents ont un lexique avec une terminologie standard. »
« *De l'autre, au niveau de l'ingénierie des exigences des SI, on a des langages et des extensions de langages existants permettant de gérer les risques.* »
« *Le problème est que la terminologie utilisée est différente selon le langage et les standards.* »

Nous nous trouvons donc face à un domaine au vocabulaire plus ou moins établi, mais la définition de celui-ci n'est pas encore uniforme. « *Le travail consiste à essayer d'uniformiser la terminologie.* »

Parfois, le vocabulaire n'est pas encore fixé et l'on doit passer par une première étape de création de la terminologie. « *On recherche les concepts essentiels sur la gestion des risques en sécurité, c'est-à-dire les principes qui reviennent tout le temps.* »

Différentes fonctions sont attendues d'un logiciel d'analyse de texte afin de réaliser cette uniformisation.

Il doit tout d'abord pouvoir procéder à une **indexation** du texte, telle qu'elle a déjà été présentée dans l'interview de Michaël Petit, et présenter cette indexation au chercheur de sorte qu'il puisse retrouver facilement les termes à uniformiser.

L'idéal serait d'avoir une **indexation en parallèle** des différents textes à analyser afin d'avoir une **vue simultanée** d'un même concept dans les différents textes, ainsi que de ses définitions.

Si la terminologie n'est pas encore fixée et que l'on se demande quels termes devront être uniformisés, il peut être intéressant d'avoir une indication de leur fréquence (« **lexicométrie** ») dans chaque texte, même si celle-ci n'est pas toujours déterminante.
« *La fréquence peut être une indication pour déterminer quel concept est propre au domaine, mais pas toujours. Il faut que ce soit plus fréquent que dans d'autres documents.* »

Une fois le repérage des termes effectués, il s'agit de « *voir les différences de vocabulaires pour un même concept* » via un « *matching sémantique* » qui est fait manuellement.

Les différentes définitions sont comparées afin de déterminer leur identité, fonction d'**« analyse des correspondances »**. Il s'agit de voir « *à quel pourcentage les concepts sont équivalents* ».
Le logiciel pourrait indiquer les termes communs entre les définitions d'un même terme.

Nous nous retrouvons toutefois toujours face au même problème. « *Un problème vient de l'ambiguïté des définitions.* » A supposer que deux définitions soient identiques, rien ne nous permet d'affirmer qu'elles ont exactement la même sémantique pour les deux auteurs qui les ont formulées.

L'étape suivante est la construction d'un méta-modèle. « *Après avoir identifié ces concepts et leurs définitions, on peut construire un méta-modèle représentant de manière plus formelle les concepts d'analyse de risque et leurs liens.* »

Il s'agit ici de restructurer les termes de l'ontologie et les liens entre ceux-ci. Une fonctionnalité utile à ce sujet serait de pouvoir présenter cette ontologie (fonction « **présentation structurée** »).

Enfin, le méta-modèle construit est comparé avec d'autres méta-modèles. Nous retrouvons donc la fonction d'**« analyse des correspondances »**.

- Objectif : définir la sémantique ontologique d'un langage

Une deuxième attente identifiée dans le cadre de l'interview est la définition de la sémantique ontologique d'un langage.

« *L'ontologie est un concept issu de la philosophie et notamment du philosophe Mario Bunge qui a voulu identifier les concepts de base pour décrire et classifier le monde physique. Sa théorie se base sur la notion de « chose ».*
L'idée est de reprendre cela pour modéliser les concepts des SI. »

« *On prend la classification et on en fait un modèle conceptuel, on prend le méta-modèle d'un langage qu'on veut étudier, et on cherche des liens entre les deux.*
A la fin, on obtient la sémantique ontologique des concepts. »

Ce modèle devrait pouvoir être construit à l'aide d'un logiciel (« **construction du modèle** »). Il devrait ensuite être possible d'importer le méta-modèle est d'établir des liens entre les deux. Le logiciel devrait donc « *pouvoir automatiser l'analyse des correspondances. Est-ce que deux langages représentent le même chose ?* ». Nous sommes à nouveau face à la fonction d'**« analyse des correspondances ».**

Nous nous trouvons ici aussi face au problème de la « **gestion des ambiguïtés** »

« *Un outil utile devrait pouvoir repérer les termes ambigus. Et s'il y a des synonymes, il devrait pouvoir le signaler.* »

- Objectif : extraire des modèles conceptuels

Enfin, une dernière attente concerne l'extraction de modèles conceptuels à partir de textes.

« *Le problème est de comprendre la sémantique de ce qu'il y a dans le texte* ».

Il existe actuellement des tentatives basées sur l'« *utilisation de langages naturels contrôlés avec l'imposition d'une syntaxe restreinte pour les descriptions* ». Si le logiciel sait quels termes il doit trouver, il est plus facile d'automatiser l'analyse et l'extraction du modèle.

3. Analyse de l'interview de Anne Devos

Lors de son interview, Anne Devos nous a expliqué la méthodologie qu'elle applique lorsqu'elle est amenée à analyser des matériaux textuels.

L'objectif poursuivi est de coder le matériel et de procéder ensuite à des comparaisons.

Celui-ci est de trois types :
« *1) des interviews semi-directives ;*
2) des documents d'entreprise : journaux, procès-verbaux de réunions ;
3) des observations participantes ».

Il convient de noter au préalable que « *Tous les textes sont récoltés par rapport à un objet précis en lien avec une recherche. On travaille dans un cadre conceptuel théorique qui oriente le regard et les questions posées dans les entretiens.* »

Avant toute analyse, Anne Devos établit une grille de codage sur base du cadre conceptuel théorique de la recherche, « *où sont définies les principales thématiques en fonction de ce qui est recherché* », chaque thème se voyant attribuer un code.

Une première attente serait donc la « **gestion d'une grille de codage** », celle-ci présentant différents aspects.

Un premier aspect est la construction physique et sémantique de la grille en elle-même (« *aider à la construction de la grille de codage* »). Construction physique dans le sens d'organisation des différents thèmes entre eux et de présentation de cette organisation.
Construction sémantique dans le sens d'identification des concepts de la grille.
Nous nous trouvons donc déjà face à deux fonctions possibles pour un logiciel : « **construction physique** » et « **construction sémantique** ».

La construction physique suppose de pouvoir créer une grille à partir de rien et de montrer comment les différents éléments s'articulent.

La construction sémantique suppose une analyse du cadre conceptuel théorique de la recherche. Le logiciel devrait identifier les concepts clés du domaine et les proposer au chercheur pour validation.

En ce qui concerne le corpus, celui-ci est découpé en blocs par le chercheur. Il pourrait donc être utile d'avoir un logiciel reprenant cette découpe (fonction « **présentation du texte** »).

Une fois la grille établie, celle-ci est appliquée au texte. « *Un logiciel pourrait nous aider à appliquer la grille au texte en transformant les codes en mots-clefs et en indexant le texte.* » Il s'agirait pour le logiciel de déterminer si un des codes est applicable aux blocs analysés

(fonction « **thématisation** »). Cela exige une analyse du sens. « *Parfois, le sens est implicite. Une automatisation complète n'est donc pas possible.* »
Cette recherche du sens n'est pas quantitative. « *On n'utilise pas d'analyses statistiques.* »

« *ce serait bien d'avoir un logiciel qui affiche le texte et la grille en parallèle.* » (« **affichage parallèle** »).

Ce travail d'application est produit de manière itérative. Trois ou quatre passages sont effectués au cours desquels la grille croît puis décroît. Les codes sont affinés puis regroupés. L'idée étant qu'au final, il y a un seul code par bloc.
« *La grille de codage basée sur le cadre conceptuel et les questions devient descriptive, puis on opère des regroupements quand les codes sont proches.* »

A ce sujet, « *Un logiciel devrait pouvoir être nourri et permettre d'organiser clairement les codes.* » On en revient à la fonction de « construction physique » qui est en perpétuelle évolution. La grille doit pouvoir être enrichie (fonction « **modification** » de la grille).
De plus, il est intéressant de garder une trace de ces modifications. Il est « *intéressant de pouvoir garder des liens entre les subdivisions* » (fonction « **historique** »).

Le travail d'analyse et de codage passe également par une phase d'identification de « *citations exemplatives* » (« **Recherche de phrases clés** »).

Une fois le travail de codage effectué, le chercheur procède à des comparaisons. « *On cherche les rapports entre le fonctionnement de l'organisation et la question de recherche.*
On fait ces comparaisons au sein de groupes d'acteurs et entre groupes, et même entre études. »
En ce qui concerne la méthodologie employée, « *on recherche des régularités entre des idées et des oppositions, plus que des fréquences* ».

Une aide logicielle pourrait consister en un « **affichage simultané** » des différents blocs ayant le même code. Le chercheur pourrait ainsi plus facilement procéder aux comparaisons.

4. Synthèse des interviews

Les recherches effectuées pour le premier chapitre de ce mémoire nous ont permis de découvrir les différents objectifs qui peuvent présider à une « analyse de texte ». Comme nous l'avons déjà dit, cette discipline semble avoir été fortement investie par les domaines des sciences humaines, sociales, économiques et littéraires, et beaucoup moins par des domaines plus scientifiques et techniques.

L'analyse des interviews nous montre que l'analyse de texte n'est pas absente des secteurs plus techniques, mais qu'elle poursuit des objectifs différents, et demandent donc des moyens différents.

Elle nous montre également la diversité des attentes en matière d'analyse de texte. Nous allons cependant essayer d'en faire une brève synthèse, mais également le tri.

En ce qui concerne les objectifs poursuivis, certains ressortent clairement des logiciels d'analyse de texte présentés dans le chapitre 3, et d'autres non.

Toutes les demandes concernant une ontologie (création, structuration, uniformisation, etc.) et celles portant sur l'application ou l'extraction d'un modèle (Messieurs Petit et Heymans) nous ont semblé ressembler aux analyses de texte auxquelles nous sommes confrontés depuis le début de ce mémoire et en particulier à celles offertes par les méthodes informatisées. Nous nous centrerons don sur celles-ci par la suite.

L'activité d'analyse de texte dans ces cas demande principalement des outils orientés sur l'analyse des mots d'un texte : indexation (éventuellement de plusieurs textes en parallèle), gestion des ambiguïtés, recherche de connexité entre termes (synonymes), comptage, contraste, recherche de correspondances, ainsi que des outils de présentations : affichage des contextes, en parallèle, des définitions, etc.

Les autres demandes (recherche de site, organisation de documents, etc.) s'éloignent par contre de l'objet de ce mémoire.

En ce qui concerne l'activité de Madame Devos, nous sommes en plein dans l'analyse de contenu telle que cette notion a pu être abordée dans le chapitre 1.

Son activité d'analyse est centrée sur des blocs de texte qu'il s'agit de coder en fonction d'une grille préalable. Sont donc plus utiles des outils permettant de construire cette grille de l'appliquer et d'en afficher les résultats.

Nous allons donc maintenant passer au dernier chapitre de ce mémoire et déterminer si les logiciels retenus présentent une utilité pour le monde de la recherche.

Chapitre 5 : Confrontation des exigences aux logiciels

Dans le cadre de ce dernier chapitre, nous allons confronter les exigences des chercheurs en matière d'analyse de texte aux fonctionnalités des logiciels présentés dans le chapitre 3.

Pour rappel, nous avons pu dégager trois grandes catégories d'objectifs ressortant de l'analyse de texte, deux dans le secteur de l'informatique (ontologie et modélisation) et un dans les sciences humaines (gestion et application de grilles de codage sémantiques).

L'ontologie vise à définir le vocabulaire clé d'un domaine.

Les attentes concernant ce point sont larges : création, structuration, uniformisation, etc. Les fonctionnalités identifiées étaient les suivantes :
- Indexation (en parallèle) ;
- Gestion des ambiguïtés ;
- Affichage du contexte (en parallèle) ;
- Indication des fréquences ;
- Analyse des correspondances/ comparaison ;
- Contraste du vocabulaire entre textes ;
- Affichage des définitions ;
- Présentation de l'ontologie ;
- Vérification de la pertinence d'autres textes.

La modélisation a pour objectif de représenter les éléments essentiels d'un existant sous forme de diagrammes et schémas. Elle a une visée d'abstraction.

Dans ce cadre, un logiciel devrait pouvoir aider à sa création et à son application. La fonction suivante est apparue nécessaire :
- recherche de termes ;
- analyse des correspondances / comparaison.

Enfin, la construction et l'application de grilles de codage demandent les outils suivants :
- Construction physique de la grille ;
- Construction sémantique de la grille ;
- Gestion de la grille : modification – historique ;
- Présentation du texte ;
- Affichage parallèle de la grille et du texte ;
- Thématisation ;
- Identification des phrases clés ;
- Affichage simultané de blocs de texte ayant le même code.

Nous allons maintenant confronter ces différentes demandes aux logiciels présentés.

1. Confrontation des objectifs en matière informatique

Comme nous l'avons vu, les exigences des chercheurs informaticiens en terme d'analyse de texte sont particulières. Leurs objectifs ont trait à tout ce qui tourne autour, soit d'une ontologie (création, mise en forme, uniformisation, comparaison, etc.), soit de l'application d'un modèle existant (recherche de concepts pouvant ressortir du modèle).

Dans le cadre de ces analyses, l'unité de base est le terme. L'ontologie a pour objectif d'identifier le vocabulaire clé d'une discipline ainsi que ses relations intrinsèques. L'application d'un modèle existant se base sur la recherche de termes pouvant refléter un élément du modèle.
Ce n'est éventuellement que dans un second temps que l'analyse portera sur des unités plus larges (phrases, paragraphes). Par exemple, pour procéder à des comparaisons de définitions de termes dans le cadre de l'ontologie, ou pour pouvoir affiner la construction du modèle.

Il apparaît donc que les logiciels les plus à même d'apporter une aide aux chercheurs informaticiens sont ceux se basant sur la même unité d'analyse, à savoir les logiciels lexicométriques au sens large tels AntConc, Lexico ou Unitex que nous avons approfondis, mais également d'autres figurant dans la liste en annexe.

Nous allons maintenant passer en revue les attentes en terme d'outils.

Une capacité attendue d'un logiciel est de pouvoir indexer le corpus, c'est-à-dire de lister tous les termes qui s'y trouvent.

Les trois logiciels lexicométriques, AntConc, Lexico et Unitex, offrent cet outil. Il convient toutefois de rappeler que les méthodes utilisées sont différentes. Lexico se base sur le repérage de délimiteurs et Unitex sur des délimiteurs, des dictionnaires et des grammaires. Les résultats seront donc différents et il conviendra de les analyser avec prudence.

Lexico n'offre donc pas de lemmatisation des termes. Les singuliers, pluriels, féminins, masculins, formes conjuguées des verbes ne seront donc pas regroupés. Par contre, les homonymes le seront.

AntConc offre de plus la possibilité d'indexer plusieurs textes en même temps.

Cet outil d'indexation nous semble essentiel en terme de support à l'analyse. Son principal intérêt est de présenter le corpus à analyser sous un autre format qui facilite le repérage des termes significatifs pour le chercheur puisque débarrassés de leur contexte qui peut parasiter leur indentification.
Une présentation sous forme de liste des termes du corpus permet de passer plus facilement et rapidement ceux-ci en revue et de cerner ceux qui sont à retenir. L'identification des synonymes s'en trouve également simplifiée, puisque chaque terme peut être envisagé pour lui-même.

Il convient de noter que Tropes effectue également une indexation mais partielle. Seuls les substantifs les plus significatifs sont retenus, sur base de choix théoriques des concepteurs. Il est donc possible que le logiciel passe à côté de termes importants pour le chercheur, les choix des uns et des autres pouvant ne pas converger. De même, chaque terme est classé dans un univers de référence et ce classement peut également apporter des surprises.

Sémato, par contre, n'offre pas d'outil de ce genre. Il effectue une indexation interne, basée sur la lemmatisation, mais n'en affiche pas les résultats.

En lien avec l'indexation a été soulevé le problème des mots ambigus.
Ni AntConc, ni Lexico ne gèrent ce problème. Toutes les formes graphiques identiques seront rassemblées (exemple : « avions » du verbe « avoir » et du substantif « avion » apparaîtront

ensemble). Il appartiendra donc d'être attentif lors de l'examen des résultats afin d'isoler les éléments significatifs.

Unitex, lui, utilise également les délimiteurs, mais est également équipé de dictionnaires reprenant toutes les possibilités pour une forme et de grammaires décrivant les différents contextes où peuvent apparaître les limites de phrases. Ces différents outils, appliqués automatiquement lors du prétraitement du texte, sont censés lui permettre de gérer un maximum d'ambiguïtés. De plus, Unitex offre également d'autres outils de levée des ambiguïtés (automate du texte, construction de grammaires de levée des ambiguïtés). Mais ces outils sont très complexes à utiliser.

Tropes prétend lui aussi gérer les ambiguïtés en se basant sur des algorithmes d'intelligence artificielle. Selon les concepteurs, « *Il lui est impossible d'effectuer ce travail de façon parfaite, mais son taux d'erreur est suffisamment faible pour permettre une analyse correcte de votre texte* ».[172]

Lexico, Unitex et AntConc sont également équipés d'un outil de concordance qui permet d'afficher les contextes d'un terme.
Cet outil présente son utilité. Dans le cadre de l'uniformisation d'une ontologie, il est utile de pouvoir afficher simultanément les phrases contenant un certain terme afin de comparer les définitions qu'elles pourraient contenir. Plutôt que d'utiliser un traitement de texte habituel qui demande de copier-coller tous les contextes d'un terme, il apparaît plus rapide d'utiliser ce genre de logiciel qui effectue l'opération quasi instantanément. Le gain est donc essentiellement en terme de rapidité et de présentation visuelle. Il en est de même pour l'application d'un modèle. Cet outil permet de visualiser toutes les parties de textes comprenant un terme, élément du modèle.

Tropes permet également d'afficher les contextes des termes figurant dans les différentes listes produites (univers de références, références utilisées).

Sémato, via l'outil requête, permet également d'identifier et d'afficher tous les textes comprenant un terme ou une expression précise.

L'indexation s'accompagne généralement d'une indication de la fréquence des termes. C'est le cas pour AntConc, Lexico et Unitex. Si ce critère n'est pas déterminant pour l'identification des termes, il peut parfois être un indicateur parmi d'autres. Tropes affiche également les fréquences des catégories de termes retenues (univers de référence, références utilisées). Quant à Sémato, il n'affiche pas la fréquence des termes, mais via les requêtes, il est possible d'obtenir la fréquence de phrases et de textes comprenant un terme ou une expression précise.

Une autre demande concernait l'analyse des correspondances, c'est-à-dire la comparaison de définitions. Il s'agit d'une opération complexe. Quand deux définitions peuvent-elles être considérées comme similaires ? Par l'emploi des mêmes termes ? Même dans ce cas, rien ne dit que la configuration des phrases ne sera pas telle qu'elles ne diront pas la même chose ? De plus, comment gérer les synonymes ?
Aucun logiciel n'offre d'outil pour gérer ce problème. Tout au plus, Sémato, via les « réseaux de similitudes » prétend identifier les textes similaires deux à deux. Cependant, les critères qui

[172] www.acetic.fr

fondent ces similitudes n'ont pas été identifiés. Et il n'est donc pas possible de dire si cet outil apportera un résultat probant.

Après avoir construit une ontologie, il est important de pouvoir la présenter. Selon nous, Tropes offre un outil intéressant en la matière : les scénarios. Ceux-ci permettent de présenter l'ontologie sous forme d'arborescence. Neufs niveaux de profondeur sont possibles. Toutefois, tous les liens entre termes ne pourront pas être indiqués puisque des cycles ne sont pas possibles.

Ces scénarios permettront d'aider le chercheur dans deux autres aspects de la gestion d'une ontologie : la vérification du contraste de vocabulaire entre textes et la vérification de la pertinence d'autres textes.

En effet, il est possible d'appliquer le scénario ainsi construit à d'autres corpus du domaine afin de déterminer la valeur de l'ontologie établie. Appliqué à des textes d'autres disciplines, il permettra de voir si les termes retenus y sont présents ou absents. Appliqués à d'autres corpus du domaine, il permettra de vérifier leur pertinence en indiquant si les termes retenus y sont présents.

En conclusion, l'aide qu'un logiciel d'analyse de texte peut apporter nous semble donc essentiellement pratique : recherche de termes précis, affichages de leur contexte, indication des fréquences, construction de l'ontologie et application à d'autres corpus.
Les trois logiciels lexicométriques offrent des outils comparables sur ces points. AntConc offre toutefois l'avantage de permettre l'utilisation de ces outils sur plusieurs textes à la fois.
Tout ce qui a trait à l'analyse au sens strict et à l'interprétation (comparaison de définitions, liens entre les termes) reste de la compétence du chercheur.

2. Confrontation des objectifs pour les sciences humaines

L'analyse de l'interview de Madame Devos nous l'a montré, ses attentes en matière d'analyse de texte sont différentes.

Son travail consiste à coder les matériaux textuels recueillis dans le cadre d'une recherche à l'aide d'une grille d'analyse. Celle-ci est préalablement construite sur base des concepts théoriques sous-tendant la recherche et est modifiable au fur et à mesure de son application au corpus.

Les attentes en termes de logiciels d'analyse de texte sont de pouvoir aider à la conception de la grille et à son application au corpus.

L'unité d'analyse n'est plus ici le mot, mais le bloc de texte.

Il apparaît donc que les logiciels qui semblent les plus à même d'apporter une aide sont ceux de la catégorie socio-sémantique : Sémato et Tropes.

En ce qui concerne la présentation et la découpe du texte en blocs, Sémato procède automatiquement à celle-ci lors du transfert du texte sur base des retours à la ligne introduits par le chercheur.

Sémato permet également d'introduire des balises, mais dans sa version payante.

Concernant la construction de la grille d'analyse, tant Sémato que Tropes offrent un outil intéressant, mais qu'il convient de manipuler avec prudence.

Sémato, via son assistant scripteur de thèmes, permet de créer ses propres grilles thématiques. Chaque thème est identifié par des ingrédients (mots, phrases, expressions) personnalisables par l'utilisateur. Pour cette personnalisation, le chercheur peut se faire aider par un assistant, mais il ne s'agit que d'une aide dont tous les résultats doivent être questionnés et approuvés par le chercheur.

Sémato offre également un outil de génération automatique de thèmes, mais dont les résultats doivent être interrogés au regard des concepts théoriques de la recherche. Au mieux, ils peuvent être une source de questionnement pour le chercheur qui ne doit pas les accepter comme tel. Ils peuvent d'ailleurs être modifiés.

Tropes offre l'outil « scénario » qui permet de créer ses propres arborescences thématiques. Tropes est également fournis avec des scénarios préétablis, mais ceux-ci peuvent ne pas correspondre aux classifications du chercheur. Ils doivent donc être utilisés avec prudence.

Toutefois, un inconvénient de ces deux outils est qu'on en revient toujours au mot, alors que les analyses effectuées par Madame Devos ne sont pas lexicales, mais purement sémantiques. C'est la combinaison d'un ensemble d'éléments qui donne son sens à un bloc de texte, combinaison difficilement formalisable.

Nous pensons toutefois que cette difficulté peut être contournée concernant Sémato, par l'application des thèmes au corpus. En effet, cette application peut être totalement manuelle via les pages d'arrimages.
Par la suite, il est possible de rechercher les blocs de texte ayant le même thème, via l'outil « requêtes en repérage », afin d'affiner ledit thème ou de comparer les blocs.

L'application d'un scénario dans Tropes ne nous semble par contre pas contourner le problème de l'analyse sur base lexicale. En effet, Tropes recherche les phrases contenant les termes mis dans le scénario. Nous sommes donc toujours dans une perspective lexicale. Il n'est pas possible de décider que telle branche du scénario sera attachée à tel bloc du texte.

Concernant l'affichage en parallèle de la grille et du texte, seul Sémato offre cette possibilité par ses pages d'arrimages. Il est possible d'avoir un affichage par bloc ou par phrases. Tous les thèmes retenus sont affichés en parallèle et il est possible d'attacher un ou plusieurs thèmes aux éléments (blocs, phrases).

En ce qui concerne la thématisation vue comme l'application automatique des thèmes au corpus, Sémato offre la possibilité d'afficher tous les textes correspondant à un thème qu'il a lui-même généré, c'est-à-dire contenant au moins un des ingrédients de ce thème. Sémato opère donc des arrimages par lui-mêmes. L'utilisateur peut toutefois modifier ces résultats.

Concernant la gestion de la grille (historique, modification), Sémato offre la possibilité de fusionner des thèmes. Toutefois, cette fusion faite, il n'est plus possible de revenir en arrière. Il n'y a donc pas d'historique à ce sujet. Il est également toujours possible de modifier ceux-ci. Sémato garde cependant un mémoire les thèmes qui ont été attachés ou supprimés de blocs.

Via l'outil requête, il est possible d'afficher simultanément des blocs de texte ayant le même code.

Enfin, lors de son analyse, le chercheur procède également à l'identification des phrases clés. Tropes prétend faire de même en se basant sur l'analyse cognitivo-discursive.[173] Il convient donc de déterminer si la stratégie suivie par ce type d'analyse correspond aux attentes du chercheur.

En conclusion, Sémato offre différentes fonctionnalités pratique qui peuvent aider le chercheur en science humaines (organisation de sa grille de codage, affichage simultanée de la grille et du texte, affichage de blocs avec le même code). A nouveau, les fonctions plus avancées qui prétendent se substituer au chercheur dans la phase d'analyse au sens strict sont à considérer avec prudence. Les résultats produits doivent être questionnés et approuvés (ou non).

Quant à Tropes, il offre peu d'aide pratique. La plupart des fonctionnalités donnent des résultats d'analyse qui doivent être interrogés par le chercheur.

[173] « chaque proposition du texte se voit attribuée un score calculé en fonction de son poids relatif, de l'ordre d'arrivée et de son rôle argumentatif. Les propositions sont ensuite triées, puis filtrées en fonction de leur score. », http://www.acetic.fr/semantique-4.htm

Conclusion de ce chapitre

Le but de ce chapitre était de déterminer si certains logiciels d'analyse de texte pouvaient répondre aux attentes des chercheurs, qu'ils soient informaticiens ou chercheurs en sciences humaines.

Deux grandes catégories d'outils avaient été identifiées dans le chapitre trois : ceux de nature plutôt lexicométrique dont l'unité d'analyse est le mot et ceux de nature sémantique où l'unité de traitement est la phrase ou le bloc de texte.

Ce chapitre nous a montré que cette subdivision se reflétait dans les attentes des chercheurs puisque, ici aussi, les attentes recueillies dans le cadre strict de ce mémoire sont apparues différentes selon l'objet de la recherche.

Les activités d'analyse de texte des informaticiens sont d'abord axées sur le mot, unité la plus élémentaire, que ce soit dans la construction d'une ontologie ou l'application d'un modèle. Les logiciels lexicométriques sont donc plus à même d'offrir des outils correspondant à leurs besoins.

En ce qui concerne les sciences humaines, les analyses de texte visent à coder des blocs de texte en fonction du sens de ceux-ci. Des logiciels travaillant dans ce sens sont donc plus appropriés.

Toutefois, la conclusion finale qui se dégage est que certains outils peuvent apporter une aide pratique (affichage, recherche, etc.). Tout ce qui a trait à l'analyse au sens strict et à l'interprétation restera toutefois de la compétence du chercheur.

Conclusions

Le premier chapitre de ce mémoire qui était consacré à la notion d'analyse de texte nous a permis d'identifier les différentes approches conceptuelles qui ce cachent derrière cette notion. L'accent a été mis sur une approche dominante, l'analyse de contenu. Ses différentes formes ont été identifiées et classées les unes par rapport aux autres. Les buts que ces formes poursuivent et les méthodes qu'elles utilisent ont également été mis en avant. Deux grandes familles ont donc été identifiées au sein de l'analyse de contenu : les formes d'analyse qui s'intéressent au fond du texte afin d'identifier son objet et celles qui se penchent sur sa forme afin d'en analyser la structure.

Le second chapitre nous a permis de nous familiariser avec les méthodes d'analyse de texte qui ont été traduites dans des logiciels informatiques. La grande majorité de ces méthodes ont pu être rattachées à l'analyse de contenu qui se penche sur le fond du texte pour en trouver l'objet. Deux grandes catégories ont pu être dégagées. Nous avons d'une part les méthodes quantitatives qui produisent des résultats sur la base de décomptes ou de calculs statistiques, et d'autre part les méthodes qualitatives qui se penchent sur les qualités de certains éléments du texte.

Le troisième chapitre nous a fait pénétrer au cœur des logiciels d'analyses de texte. Certains ont été sélectionnés afin d'en faire une analyse approfondie, de présenter les outils qu'ils offrent et de les évaluer. Le choix des logiciels a eu pour objectif de montrer la diversité des outils proposés. Les fonctionnalités communes ont toutefois pu être comparées. Il ressort de cette présentation que certains logiciels apportent une aide pratique et laisse une grande liberté à l'utilisateur, tandis que d'autres prétendent effectuer une partie de l'analyse et appliquent automatiquement les outils qu'ils proposent.

Le chapitre quatre était consacré aux attentes des chercheurs par rapport à de tels logiciels. Des interviews ont permis de les cerner, mais également de montrer la diversité des exigences, diversité liée à l'objet qu'ils poursuivent lors d'une recherche incluant d'une analyse de texte.

Enfin, le cinquième chapitre a confronté les attentes mises en avant au chapitre quatre avec les outils identifiés dans le chapitre trois. Il ressort qu'à attentes différentes correspond des outils différents. Certains logiciels peuvent apporter une aide pratique au chercheur. Toutefois, une certaine prudence doit être de mise vis-à-vis des outils qui tentent de se substituer à l'analyse du chercheur. En effet, ils reposent sur les présupposés des concepteurs qui ne sont pas nécessairement ceux de l'utilisateur. Leurs résultats doivent donc être questionnés.

Ce mémoire, de nature exploratoire, a donc permis de dresser une cartographie des logiciels d'analyse de texte existants et d'investiguer certaines éléments de cette carte. Il a également eu pour objectif de mettre en avant les potentialités de ces outils, souvent ignorées du monde scientifique. Ce premier aperçu des forces et faiblesses pourrait être poursuivi dans un autre mémoire qui approfondirait les choix opérés et confronterait de manière plus fouillée, in vivo, les fonctionnalités offertes aux réalités des analyses.

Bibliographie

1. Articles et livres

Ambroise, C., « Classification hiérarchique », http://www.hds.utc.fr/sy09/documents/hierarchie.pdf

Audet, C.-A., « Méthode d'analyse structurale de la phrase », http://www.aide-doc.qc.ca/le.grammairien/ftp/analstruct.pdf

Bardin, L., L'analyse de contenu, 11° édition, Paris, 2003, P.U.F.

Fayol, M., Le récit et sa construction, Paris, Delachaux et Niestlé, 1985.

Ghiglione, R. et al., L'analyse automatique des contenus, Paris, Dunod, 1998

Jenny, J., « Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine. Etat des lieux et essai de classification. », Article publié dans le *Bulletin de Méthodologie Sociologique (B.M.S.),* n° 54, mars 1997, p.64-112, http://pageperso.aol.fr/jacquesjenny/ATBMS.htm

Patanella, N., « Note sur la méthodologie de la science politique - *A destination des étudiants du DEA en relations internationales et intégration européenne* », http://www.ulg.ac.be/polgereg/Publications/Methodo.pdf

Renaud, L., « Méthodes de recherche en communication », mars 2003, http://www.er.uqam.ca/nobel/r13761/medias/cours10.pdf

Reuter, Y., L'analyse du récit, Paris, Dunod, 1997.

Revollon, P. et Larrecq, C., « L'analyse de contenu », http://www.inh.fr/enseignements/idp/idp2005/outils/etude_marche/contenu_psycho_socio.pdf

Scanu, A. M., « Hyperbase – Un logiciel pour l'analyse textuelle », www.rilune.org/dese/tesinepdf/**Scanu/Scanu**_Litt%E9ratureetinformatique.pdf

Schurmans, M.-N., « Introduction aux démarches compréhensives », 12 mai 2004, http://www.unige.ch/fapse/SSE/teachers/schurmans/notesCours12%20mai.rtf

Schwischay, B., « Mémento d'analyse grammaticale », 3 mars 2002, http://www.home.uni-osnabrueck.de/bschwisc/archives/analyse.pdf

Suter, C., « L'analyse de contenu », Sociologie générale I : Introduction à la sociologie, http://www.unine.ch/socio/enseignement/socio1_2004/soc1_8_6.doc

2. Manuels

« Modalisa – Liste des fonctionnalités », http://www.modalisa.com/

« Modalisa - Première visite », http://www.modalisa.com/

« Tropes® *Version 7.0* Manuel de référence ».

Antony, L., « Read me File for AntConc 3.1.303 (Windows and Linux) » du 31 juillet 2006

Fracchiolla, B., Kuncova, A. et Maisondieu, A., « Manuel d'utilisation », http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/manuels.htm

Laprun, C., Fiscus, J. G., Garofolo, J., Pajot, S., « A Practical Introduction to ATLAS », http://www.nist.gov/speech/atlas/

Maisondieu, A., « Manuel d'utilisation abrégé (Dix premiers pas avec *Lexico3*) », http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/manuels.htm

Mueller, J.-P., « La librairie ttda: tools for textual data analysis », 5 mars 2004, http://wwwpeople.unil.ch/jean-pierre.mueller/ttda-fr.pdf

Paumier, S., « UNITEX 1.2 - MANUEL D'UTILISATION », mai 2006

Silberztein, M., « Intex », http://intex.univ-fcomte.fr/

Zhang Le, « Morphix-NLP Live CD Manual », 26 octobre 2003, http://morphix-nlp.berlios.de/

3. Présentations

Gruau, C., « Revue des logiciels d'analyse de texte », 29 avril 2004, http://www.cemef.net/fr/presentation/pagesperso/cg-promo2001/extdoc/Gruau-AnalyseQualitative.pdf

Mongeau, P., « Effectuer la cueillette des données », http://www.er.uqam.ca/nobel/r32700/Cours%20site/SITE%20%20COM7103/cueilletteetanalyse.htm

Université Pierre et Marie Curie, « Analyse de données - Ordinations et groupements », http://www.obs-vlfr.fr/Enseignement/enseignants/labat/anado/acp/presentation.html

4. Compte-rendu d'ouvrage

Egger, S., Bastian, L., Briatte, F., « L'analyse de contenu – Laurence Bardin », Compte-rendu d'ouvrage, novembre 2005, http://phnk.com/files/m2-teq-bardin-livret.pdf

5. Sites Internet

Acetic : http://www.acetic.fr/index.htm

Alceste : http://www.image-zafar.com/index_alceste.htm

Alice : http://www.alice-soft.com/html/prod_aliceserv.htm

AntConc : http://www.antlab.sci.waseda.ac.jp/

ATALA : http://www.atala.org/

ATILF : http://www.inalf.cnrs.fr/ ; http://atilf.atilf.fr/dendien/scripts/tlfiv5/showp.exe?120;s=4127846670;p=combi.htm

Atlas : http://www.nist.gov/speech/atlas/

Centre d'analyse de texte par ordinateur (UQAM) : http://www.ling.uqam.ca/ato/index.html

Cibois, P. : http://perso.orange.fr/cibois/SitePhCibois.htm

Delafosse, L. : http://perso.orange.fr/ldelafosse/Glossaire/C.htm#computationnel

Département « Maîtrise des Sciences de l'information et de la documentation » de l'Université Paris 1 – Panthéon-Sorbonne : http://mist.univ-paris1.fr/logiciel/frame.htm

Dicodunet : http://www.dicodunet.com/definitions/referencement/occurrence.htm

Gestion de la diversité : http://www.socialeurope.com/mandiv/fr/focus_group.html

Intex : http://intex.univ-fcomte.fr/

Jenny, J., http://pageperso.aol.fr/jacquesjenny/testamentscientifique.html

La Libre Belgique : www.lalibre.be

Laboratoire Document et Sciences de l'Information : http://docsi.univ-lyon1.fr/presentation.htm

Lejeune, C. : http://analyses.ishs.ulg.ac.be/logiciels/

Lettres.net : http://www.lettres.net/files/recit.html

Lexico : http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/

Modalisa : http://www.modalisa.com/

Morphix-NLP : http://morphix-nlp.berlios.de/

Mueller, J.-P. : http://wwwpeople.unil.ch/jean-pierre.mueller/

NLT : http://nltk.sourceforge.net/

Nodepad : https://sourceforge.net/projects/nodepad/

NooJ : http://nooj.matf.bg.ac.yu/ ; http://www.nooj4nlp.net/

Office québécois de la langue française :
http://www.olf.gouv.qc.ca/ressources/bibliotheque/dictionnaires/Internet/fiches/8361002.html

R : http://www.r-project.org/

SATO : http://www.ling.uqam.ca/sato/

Schwab, D.: http://www.lirmm.fr/~schwab/pmwiki/pmwiki.php?n=Recherche.Glossaire

Sémato : http://fable.ato.uqam.ca/guidexpert-ato/gea.asp

TAMSAnalyser : http://tamsys.sourceforge.net/

TELSAT : http://telsat.belspo.be/beo/fr/guide/compprin.asp?section=3.10

Tetralogie : http://atlas.irit.fr/TETRALOGIE/tetrajeu.htm

Transcriber : http://trans.sourceforge.net/en/presentation.php

Unités d'analyse des systèmes et des pratiques d'enseignement, ULg :
http://www.ulg.ac.be/pedaexpe/cours/glosaire/acp.htm

Unitex : http://www-igm.univ-mlv.fr/~unitex/

Weblettres.net : http://www.weblettres.net/sommaire.php?entree=20&rubrique=75

Weblex : http://weblex.ens-lsh.fr/wlx/

Wikipedia : http://fr.wikipedia.org/

Wordstat : http://www.provalisresearch.com/wordstat/wordstat.html

Annexes

1. Méthode suivie par un chercheur lors d'une analyse de texte[174]

L'objet de cette annexe est de décrire les différentes étapes suivies par un chercheur lorsque l'objet de sa recherche implique une analyse de texte.

La démarche peut être découpée en trois grandes étapes :

1) la préanalyse ;
2) l'exploitation du matériel ;
3) le traitement des résultats, l'inférence et l'interprétation.

La préanalyse correspond à la phase d'organisation du travail. Elle a pour objectifs « l'opérationnalisation et la systématisation des idées de départ afin d'aboutir à un schéma précis du déroulement des opérations successives, à un plan d'analyse ».[175]

Trois choses doivent être réalisées lors de cette étape :

- le choix de documents à analyser ;
- la formulation des hypothèses et des objectifs ;
- l'élaboration d'indicateurs pour l'interprétation terminale.

De plus, avant de passer à la phase d'exploitation, le matériel doit être préparé matériellement (transcription des entretiens, mise sur fiches d'articles, numérotations de documents, etc.) et formellement (alignement des énoncés par exemple).

La phase d'exploitation du matériel consiste à appliquer la méthode d'analyse choisie, que ce soit manuellement ou par ordinateur (ex. : dans la méthode lexicographique, on procède au comptage des formes).

Cette phase produit des résultats bruts que le chercheur devra analyser plus finement. On entre dans la phase de traitement des résultats, d'inférence et d'interprétation en fonction de l'objet de sa recherche.

Quant aux logiciels d'analyse de texte, il apparaît clairement qu'ils ne peuvent intervenir qu'à deux moments.
Lors du choix des documents à analyser, ils peuvent aider le chercheur à sélectionner les textes pertinents. Si la masse de documents est trop importante pour pouvoir être lue dans son intégralité, ils peuvent aider à éliminer les documents hors sujet.
Lors de la phase d'exploitation, ils peuvent aider le chercheur à appliquer la méthode choisie.

[174] L. Bardin, L'analyse de contenu, 11° édition, Paris, 2003, P.U.F., p. 125 et s. ; A. D. Robert et A. Bouillaguet, L'analyse de contenu, 2° édition, Paris, P.U.F., 2002, p. 24 et s.
[175] L. Bardin, L'analyse de contenu, 11° édition, Paris, 2003, P.U.F., p. 125 et s.

2. Exemple d'application de l'analyse propositionnelle du discours

Cet exemple est issu de Laurence Bardin, L'analyse de contenu, 11° édition, Paris, 2003, P.U.F., p. 245-246.

« *Exemple 1* : Une étude des « besoins en formation » d'EDF-GDF a suscité une enquête par entretiens auprès des agents concernés. La technique a été mise au point sur un échantillon de 11 entretiens (sur un total de 95). Les référents-noyaux repérés sont aussi bien des notions (la formation), des instances organisationnelles (EDF, mon Service), des membres de l'organisation (mon chef, nos subordonnés) ou le locuteur (Je). Mais ils peuvent se manifester
par des équivalents, par exemple :

« *nous* sommes chargés de la vérification des comptes clients»
ou « il *y* a beaucoup de travail »
ou « les choses, *ici,* se passent plutôt bien ... »

La liste des thèmes, ou « référents-noyaux », retenus, compte une vingtaine d'entrées. Elle répond à la double contrainte d'être limitée pour être maniable, et de rendre compte d'un maximum de propositions susceptibles d'être rencontrées dans les entretiens.

Liste des thèmes ou « référents-noyaux »
- Mon Service (en tant qu'unité organisationnelle) ;
- Mon Service (en tant que groupe de personnes) ;
- Les gars (les subordonnés, les gens que je dirige) ;
- Les unités supérieures (les organismes de taille supérieure à mon service, et comprenant mon service; le Centre par exemple) ;
- Les unités extérieures (organismes hiérarchiquement indépendants de mon service, un autre district, un autre service comptable ...) ;
- La formation à l'EDF-GDF;
- L'EDF-GDF;
- La clientèle;
- Notre travail (le travail de notre service, de notre équipe, ici et actuellement);
- Moi, je;
- L'employé type EDG-GDF, le stéréotype: le gars de l'EDF;
- Mes (on) chef(s), avec qui j'ai affaire;
- La direction, en tant qu'entité; le « décideur abstrait» ;
- L'avancement;
- Les salaires, la grille;
- L'ordinateur;
- L'information à l'EDF-GDF.

Une fois la grille de thèmes élaborés, on procède à la ventilation des propositions afférentes à chaque RN, en réécrivant ces propositions une à une.

Soit, par exemple, le RN « l'employé type EDF-GDF » ; elle se voit affectée des propositions :

01 : le *gars* se plaint ici souvent
02 : *On* se demande bien ce que va devenir tout ça
03 : *Il* fait ce qu'il peut pour être en phase
04 : *Il* essaie de ne pas se laisser dépasser

Le nombre de propositions par référents-noyaux, après élimination maximale, est très variable. Dans ce cas, il varie de 4 (RN: « l'information à EDF-GDF. ») à 101 ((Moi je »), le total étant de 555 propositions réparties dans 17 RN. »

3. Logiciels

ALICE[176]

1. Caractéristiques générales

ALICE est une version allégée du logiciel AC2, tous deux commercialisés par la société ISOFT.

Il est disponible sous Windows 95/98/NT/2000.

2. Domaines d'application et fonctionnalités

ALICE peut être utilisé dans des domaines très variés : marketing, finances, assurances, ressources humaines, industries, santé, etc.

Il s'agit d'un outil de data mining et d'aide à la décision.

Il couvre le cycle complet de l'analyse, depuis la préparation des données jusqu'au déploiement de modèles

Son objectif principal est de permettre de faire des prédictions via des arbres de décisions.

Il présente les caractéristiques suivantes[177] :

Données sources	
OLE DB, OLE DB pour OLAP	X
Bases de données relationnelles	X
Fichiers SPSS, SAS...	X
Fichiers texte	X
Fichiers Microsoft (Excel, Access...)	X
Outils de requêtes	X

Préparation des données	
Prévisualisation des champs calculés	X
Fonctions d'agrégation	X
Changement du type de champ	X
Assistant de discrétisation des champs symboliques	X
Import/ export des labels de champs	X
Calcul des fréquences/ restauration des modèles	X
Discrétisation manuelle	X
Assistant de discrétisation automatique intervalles égaux	X
Assistant de discrétisation automatique intervalles égaux	X

Fonctionnalité de reporting: ALICE/REPORT	
Editeur de rapport	X
Eléments prédéfinis	X
- Statistiques descriptives	X
Champs numériques	X
Champs symboliques (synthèse)	X
Champs symboliques (détail)	X
Champs symboliques (plus fréquents)	X
Champs les plus différents	X
- Tableaux croisés	X
- Arbre	X
Prochaine coupe	X
- Graphiques	X
Camemberts	X
Histogrammes (numériques)	X
Histogrammes (symboliques)	X
- Segmentations	X
Segments supérieurs à un seuil	X
Meilleurs segments	X
Personnalisées	X
Evolution du comportement	X
Coupure/répartition du nœud sélectionné	X

[176] http://www.alice-soft.com/html/prod_aliceserv.htm ; http://www.isoft.fr/html/prod_alice.htm
[177] Tableau repris de http://www.alice-soft.com/html/prod_alice.htm

Création à la volée de champs	X
Suppression de champs	X
Rafraîchissement des données	X
Gestion des dates et des durées	X
Modification possible des formules de discrétisation ou de regroupement	X
Remplace les valeurs manquantes	X

Analyse : Arbres de Décisions	
Construction automatique	X
Construction interactive	X
Elagage	X
Représentation dynamique de l'impact d'un champ sur un nœud	X
Réduction / développement d'un niveau	X
Choix de la variable pour le développement manuel d'un nœud	X
Développement automatique des nœuds	X
Elagage, regroupement, dissociation, isolation des branches	X
Regroupement/dissociation des valeurs symboliques	X
Masquage de nœuds	X
Mise en couleur des nœuds par seuil	X
Gestion de projets à arbres multiples	X
Segmentation des variables	X
Navigation dans l'arbre	X

Convivialité	
Assistant accès aux données	X
Modèles de champs calculés	X
Assistant de discrétisation des variables numériques	X
Assistant tableaux croisés	X
Assistant graphique	X
Modèles de formules	X
Modèles de rapport	X
Boîtes de dialogue	X
Utilisation intensive des menus contextuels et du glisser-déposer	X

Fonctionnalités OLAP	
ALICE/CROSSTAB	X
- Création par glisser-déposer de tableaux croisés	X
- Mise à jour dynamique des tableaux croisés	X
Formules de nœud	X
Mise en couleur des nœuds par formules	X
Affichage de formules multiples dans le	X

Export du rapport au format RTF	X
Accessible par n'importe quel outil bureautique	X
Possibilités de copier/coller	X

Représentation graphique	
Graphiques en ligne	X
- Histogrammes	X
- Nuages de points	X
- Fréquences	X
- Courbes de densité	X
Graphiques en 2D et 3D jusqu'à 4 variables	X

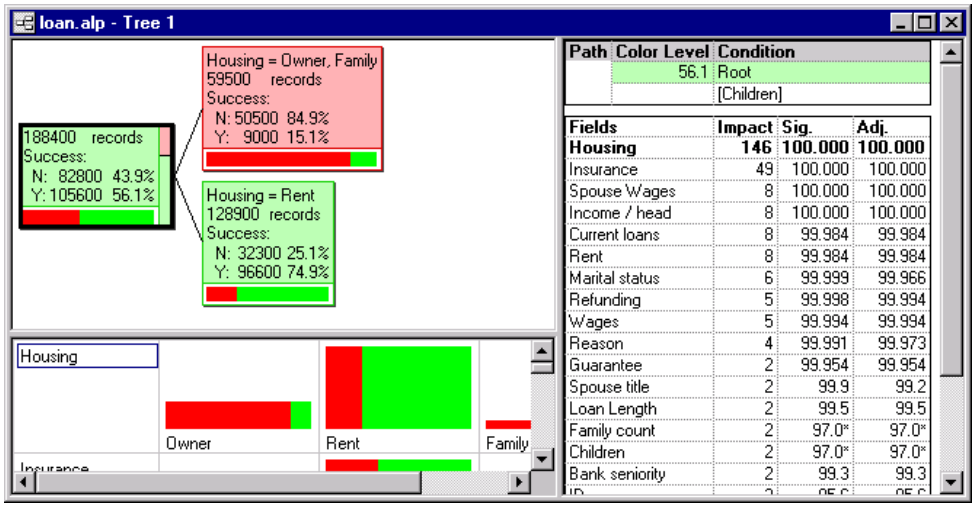
Exploitation du modèle	
ALICE/SCORING	X
- Matrice de Confusion	X
- Importation ou exportation de bases de donnés et de fichiers	X
- Prédiction de résultats sur une nouvelle population	X
- Calcul d'efficacité du modèle	X
ALICE/SEGMENT	X
- Injection d'un nouveau jeu de données	X
- Comparaison des segmentations	X
Sauvegarde de la distribution de variables	X

Fonctionnalités statistiques	
ALICE/STAT	X
- Statistiques variables numériques	X
- Statistiques variables symboliques	X
- Comparaison à-la-volée de deux segmentations	X
ALICE/CORRELATION	X
ALICE/CLUSTERING	X

Ouverture fonctionnelle d'ALICE d'ISoft	
Requêtes SQL	X
Génération du code SQL du modèle	X
Simulations "What-if"	X
Export des jeux de données et des sous-populations vers des fichiers et le presse-papier	X
Export de la plupart des éléments vers le presse-papier ou des applications bureautiques	X

nœud	

Voici un exemple d'écran type du logiciel[178] :



AntConc[179]

1. Caractéristiques générales

AntConc est un logiciel libre développé par Laurence Anthony.

Il est disponible pour Windows, MacOS et Linux.

[178] Repris du site http://www.alice-soft.com/html/prod_alice.htm
[179] http://www.antlab.sci.waseda.ac.jp/

Il est compatible Unicode et peut donc traiter toutes les langues européennes et asiatiques.

2. Domaines d'application et fonctionnalités

AntConc est un *concordanceur*, c'est-à-dire « un programme qui, pour un mot donné, recherche dans un texte toutes ses concordances, c'est-à-dire les phrases ou les groupes de mots dans lesquels il apparaît »[180].

Il permet également la génération de clusters, de N-grammes, de collocations, de listes de mots, de listes de mots clés par comparaison avec un corpus de référence, etc.

Ces différents résultats peuvent être ordonnés selon différents critères. Il est aussi possible à l'utilisateur de personnaliser ces fonctions.

Atlas (Architecture and Tools for Linguistic Analysis Systems)[181]

1. Caractéristiques générales

Atlas est issu de la collaboration entre LDC[182], MITRE et NIST[183] et est une application des travaux de Bird et Liberman sur les graphes d'annotations.

Ce logiciel ne doit pas être confondu avec le CAQDAS Atlas-ti et la station bibliométrique Atlas dont fait partie le logiciel Tetralogie.

2. Domaines d'application et fonctionnalités

Le logiciel Atlas peut être utilisé pour des corpus de nature variée : audio, écrit, image, etc.

Il a pour objectif d'organiser les annotations de corpus, indépendamment de leurs formats.

Le concept clé d'Atlas est la notion d'« annotation » en linguistique.

« *An annotation is the fundamental act of associating some content to a region in a signal.* »[184]

Un signal est défini comme : « an immutable, N-dimensional space containing phenomena that might be the target of annotations ».[185] Il s'agit d'une notion logique et non physique.

Une région est « an abstraction for identifying an area of the signal space ».[186]

[180] http://www.profession-traducteur.net/outils/outils.htm
[181] http://www.nist.gov/speech/atlas/ ; http://analyses.ishs.ulg.ac.be/logiciels/opencaqdas.html
[182] Linguistic Data Consortium
[183] National Institute of Standards and Technology
[184] C. Laprun, J. G. Fiscus, J. Garofolo, S. Pajot, « A Practical Introduction to ATLAS », http://www.nist.gov/speech/atlas/
[185] C. Laprun, J. G. Fiscus, J. Garofolo, S. Pajot, « A Practical Introduction to ATLAS », http://www.nist.gov/speech/atlas/
[186] C. Laprun, J. G. Fiscus, J. Garofolo, S. Pajot, « A Practical Introduction to ATLAS », http://www.nist.gov/speech/atlas/

Le contenu, quant à lui, représente « information that annotators would like to specify about the linguistic event occurring in the specified region ».[187]

Le processus d'annotation comprend trois étapes :

- Identification des régions intéressantes dans un signal ;
- Association d'un contenu à ces régions pour former des annotations ;
- Lier ensemble les annotations en relation.

Intex[188]

1. Caractéristiques générales

Le logiciel Intex a été développé par Max Silberztein.

Il est capable de traiter en temps réel plusieurs centaines de mégaoctets et est disponible sous Windows.

Il est capable de traiter toute langue dont l'alphabet peut être traité sur 8 bits. Dans le futur, il utilisera le codage Unicode.

2. Domaines d'application et fonctionnalités

Intex est défini comme « *un environnement de développement utilisé pour construire des descriptions formalisées à large couverture des langues naturelles, et les appliquer à des textes de taille importante en temps réel* ».[189]

Il inclut les fonctions suivantes :

- indications de statistiques sur le corpus étudié dès l'ouverture de celui-ci (nombre de phrases, de lexèmes, de formes simples, etc.) ;
- analyse du vocabulaire (reconnaissance des morphèmes[190], des mots simples, des mots composés et des expressions figées[191]) par consultation de dictionnaires ou par reconnaissance de graphes lexicaux ;
- analyse syntaxique sur base de l'analyse lexicale ;
- indexation des concordances lemmatisées ;
- mécanismes de levée des ambiguïtés ;
- fonctions de recherches par utilisation d'expressions rationnelles ;
- étude statistique des résultats ;
- éditeur graphique permettant de construire des automates à état fini (grammaires) ;
- transducteurs permettant de faire des remplacements ou des insertions dans le texte.

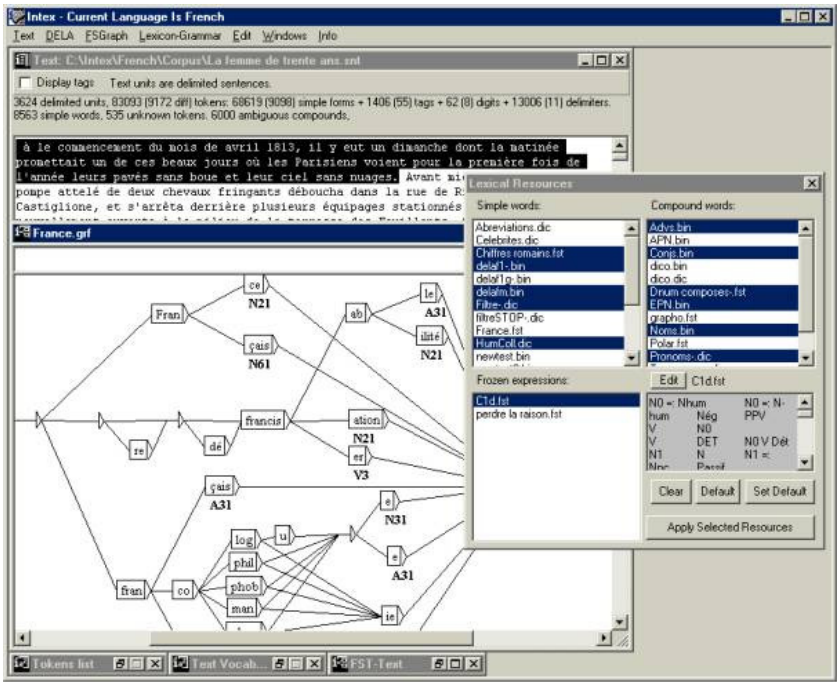
« *Une caractéristique essentielle d'INTEX est que tous les objets traités (textes, dictionnaires, grammaires) sont à un moment ou à un autre représentés par des* ***transducteurs à états finis***[192]. »[193]

---

[187] C. Laprun, J. G. Fiscus, J. Garofolo, S. Pajot, « A Practical Introduction to ATLAS », http://www.nist.gov/speech/atlas/
[188] http://intex.univ-fcomte.fr/ ; http://analyses.ishs.ulg.ac.be/logiciels/intex.html ; http://nooj.matf.bg.ac.yu/ ; Max Silberztein, « Intex »,
[189] M. Silberztein, « Intex », http://intex.univ-fcomte.fr/
[190] « des séquences de lettres inclues dans des formes simples, et associées à des informations linguistiques dans des graphes (morphologiques) », M. Silberztein, « Intex », http://intex.univ-fcomte.fr/
[191] « des séquences éventuellement discontinues de formes simples qui correspondent à des entrées lexicales dans une grammaire lexicale », M. Silberztein, « Intex », http://intex.univ-fcomte.fr/

Le travail d'indexation se fait sur base de dictionnaires fournis avec le logiciel, mais personnalisables par l'utilisateur. Dans ces dictionnaires, chaque forme est associée à son lemme et à sa catégorie morpho-syntaxique (ex. : adjectif, adverbe, substantif, etc.). D'autres informations sont parfois disponibles (ex. : code syntaxico-sémantique pour les verbes).

Exemple d'écran[194] :



Différentes ressources complémentaires (dictionnaires, outils d'analyse statistique, etc.) sont disponibles.

Lexico3[195]

1. Caractéristiques générales

Le logiciel Lexico3 a été développé à Saint Cloud par André Salem.

Il fonctionne sous Windows 95 et postérieurs, et Windows NT 3.51 et 4.0.

2. Domaines d'application et fonctionnalités

Lexico3 est un outil de statistiques textuelles.

« L'originalité principale de la série ***Lexico*** est qu'elle permet à l'utilisateur de garder la maîtrise sur l'ensemble des processus lexicométriques depuis la segmentation initiale jusqu'à

---

[192] « Un transducteur à état fini est un graphe qui représente un ensemble de séquences en entrée, et leur associe des séquences produites en sortie. », M. Silberztein, « Intex », http://intex.univ-fcomte.fr/

[193] M. Silberztein, « Intex », http://intex.univ-fcomte.fr/

[194] repris du site http://msh.univ-fcomte.fr/intex/overview.html

[195] http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/ ; http://analyses.ishs.ulg.ac.be/logiciels/lexico.html ; A. Kuncova, A. Maisondieu, « Manuel d'utilisation abrégé (Dix premiers pas avec *Lexico3*) », http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/manuels.htm ; B. Fracchiolla, A. Kuncova, A. Maisondieu, « Manuel d'utilisation », http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/manuels.htm ;

l'édition des résultats finaux. Les unités qui seront ensuite automatiquement décomptées sont exclusivement constituées à partir de la liste des délimiteurs fournie par l'utilisateur, sans recours à des ressources dictionnairiques extérieures. »[196]

Il permet le repérage des formes graphiques et l'étude de la répartition d'unités plus complexes telles les segments répétés, les types généralisés[197].

Il offre les fonctionnalités suivantes :

- segmentation du corpus ;
- concordances : visualisation de toutes les occurrences d'une *forme* ou d'un *type généralisé* (Tgen) en contexte ;
- décomptes portant sur les formes graphiques ;
- spécificités[198] et analyses factorielles portant sur les formes, les groupes de formes[199] et les segments répétés ;
- outils d'analyse : méthodes qui vont de la description statistique élémentaire (comptages, histogrammes, etc.) à divers types d'analyse multidimensionnelle des données textuelles (analyse factorielle des correspondances, classification automatique, analyse des séries textuelles chronologiques).

Il permet également une caractérisation des différentes parties d'un corpus par les formes qu'elles emploient abondamment.

Certaines méthodes ne sont pas inclues actuellement dans Lexico3 : la classification ascendante hiérarchique et les méthodes permettant de mettre en évidence les réseaux de cooccurrences.

Modalisa[200]

1. Caractéristiques générales

Le logiciel Modalisa a été conçu à partir des travaux de Philippe Cibois.

Il est disponible sous Windows et MacOS X (jusqu'à la version 4.6).

Le nombre de données pouvant être traitées est illimité.

2. Domaines d'application et fonctionnalités

Modalisa est un logiciel de création et d'analyse de questionnaires et d'entretiens, ainsi que de data mining.

---

[196] B. Fracchiolla, A. Kuncova, A. Maisondieu, « Manuel d'utilisation », http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/manuels.htm

[197] « unités de dépouillement définies par l'utilisateur à l'aide d'outils lui permettant d'effectuer automatiquement des regroupements d'occurrences du texte », B. Fracchiolla, A. Kuncova, A. Maisondieu, « Manuel d'utilisation », http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/manuels.htm

[198] « permet un jugement sur la fréquence de chacune des unités textuelles dans chacune des parties du corpus », c'est-à-dire de déterminer le sur- ou sous-emploi de l'unité textuelle, B. Fracchiolla, A. Kuncova, A. Maisondieu, « Manuel d'utilisation », http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/manuels.htm

[199] « constituer des *types* rassemblant les occurrences de formes graphiques différentes liées par une propriété commune », par ex.,rassembler le singulier et le pluriel, B. Fracchiolla, A. Kuncova, A. Maisondieu, « Manuel d'utilisation », http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/manuels.htm

[200] http://www.modalisa.com/ ; http://analyses.ishs.ulg.ac.be/logiciels/modalisa.html ; « Modalisa - Première visite » ; « Modalisa – Liste des fonctionnalités », ;

Il propose également tous les outils nécessaires à la publication de questionnaires sur Internet ou en Intranet.

Modalisa permet d'analyser des entretiens par le principe de la codification. Celle-ci consiste à sélectionner des portions de textes – unités de sens – sur lesquelles on affecte une ou plusieurs catégories. Il est ensuite possible de procéder à divers traitements, analyses et recherches, notamment l'édition des textes codifiés par catégories.
« *On passe ainsi d'un discours individuel à un discours transversal structuré en thèmes plus ou moins larges. Ces thèmes constituent eux-même une variable à réponses multiples qu'il est possible de croiser avec les variables signalétiques des personnes interviewées.* »[201]

Le logiciel permet différentes analyses lexicales :
- Liste des occurrences de mots ou d'expressions ;
- Possibilité de marquage des termes du lexique inventorié ;
- Édition d'un inventaire de toutes les portions de textes contenant au moins une des expressions marquées ;
- Regroupement de termes sur sélection dans la liste ;
- Suppression de termes dans la liste ;
- Sélection d'expression pour recherche du contexte des termes marqués ;
- Possibilité de fixer l'amplitude du contexte : nombre de caractères précédents et suivants l'expression sélectionnée.

Il offre également des fonctions statistiques, de croisement entre variables, de tris (à plat et croisés), d'analyse des cooccurrences et d'analyse factorielle des correspondances.

Modalisa est capable d'exporter des données provenant de certaines bases de données (Access, Oracle, Sybase, Lotus Notes, dBase, etc.) afin de les analyser.

Morphix-NLP[202]

1. Caractéristiques générales

Morphix-NLP est une distribution Linux compilée par Zhang Le, disponible sur un CD bootable.

Il est utilisable sur la plupart des PC de type x86.

2. Domaines d'application et fonctionnalités

Morphix-NLP contient une kyrielle d'outils libres de traitement des langues naturelles et d'exploration des contenus.

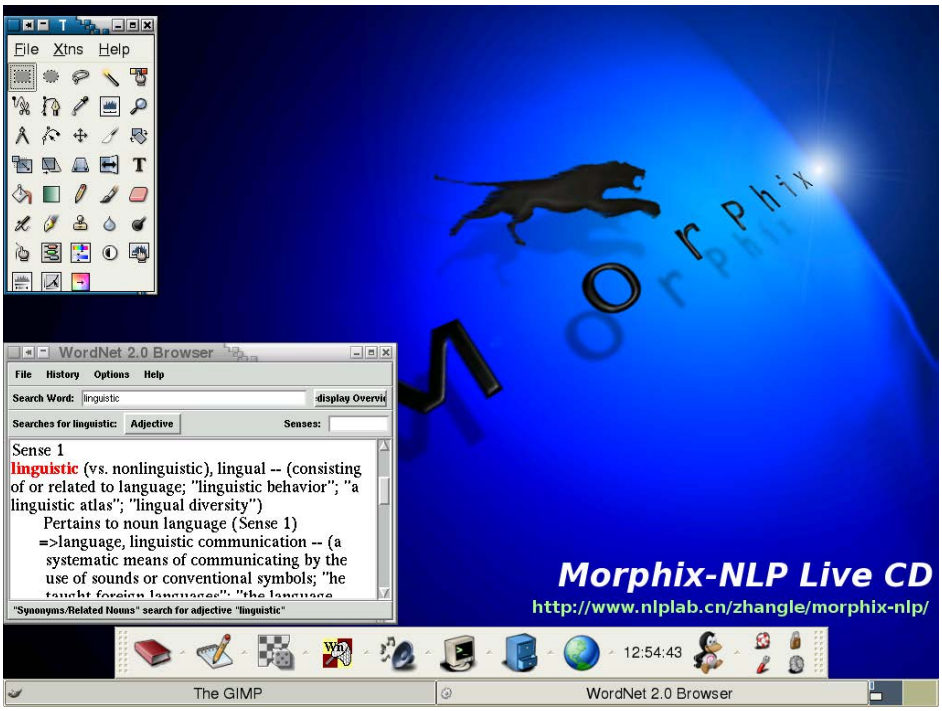
Les outils proposés sont les suivants :
- Des lemmatiseurs (tokeniser) ;
- Un système d'analyse lexicale en chinois ;
- Des « délimiteurs de parties du discours » (Part-of-Speech Tagger) ;
- Des analyseurs syntaxiques (parsers) ;

[201] http://www.modalisa.com/
[202] http://morphix-nlp.berlios.de/ ; Zhang Le, « Morphix-NLP Live CD Manual", 26 octobre 2003, http://morphix-nlp.berlios.de/

- Des outils de « Statistical Language Modeling » ;
- Des outils de développement NLP ;
- Etc.

Il contient entre autre AntConc présenté ci-dessus.

Ecran principal[203] :



Natural Language Toolkit (NLT)[204]

1. Caractéristiques générales

NLT fait partie des logiciels libres. Il a été développé par Steven Bird et Edward Loper.

2. Domaines d'application et fonctionnalités

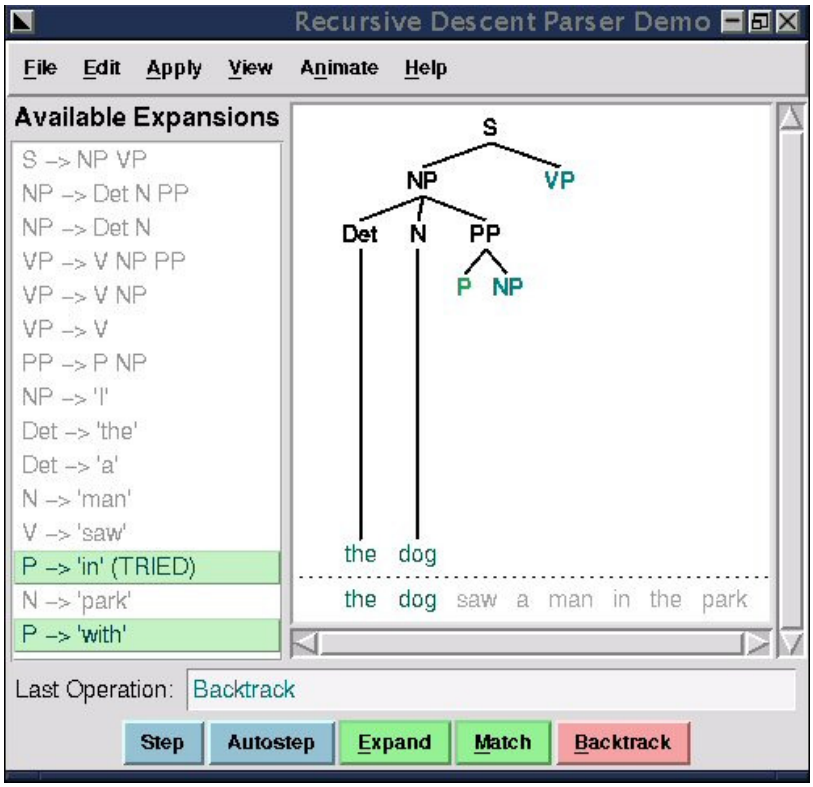
NLT offre des ressources pour la linguistique computationnelle (ou linguistique par ordinateur).[205]

NLT "is a suite of program modules, data sets and tutorials supporting research and teaching in computational linguistics and natural language processing"[206].

Il comprend les modules suivants : lecteurs de corpus, lemmatiseurs, outils d'annotations, analyseurs syntaxiques, outils de clustering, outils statistiques, etc.

Exemple d'écran[207] :

[203] Repris du site http://morphix-nlp.berlios.de/screenshot.html
[204] http://nltk.sourceforge.net/
[205] http://perso.orange.fr/ldelafosse/Glossaire/C.htm#computationnel
[206] http://nltk.sourceforge.net/
[207] repris du site http://nltk.sourceforge.net/



Nodepad[208]

1. Caractéristiques générales

Nodepad est un logiciel libre.

Il est disponible sous Windows, Posix et Linux, uniquement en anglais.

2. Domaines d'application et fonctionnalités

Nodepad est un outil d'analyse qualitative de données textuelles.

Il permet la construction d'une classification hiérarchique indexant les segments de texte d'un corpus.

Il est basé sur les idées de la « Grounded Theory methodology » quant à la catégorisation et à la conceptualisation des données dans la recherche qualitative.

Cette méthode peut être résumée comme suit : « *In short, in using grounded theory methodology you assume that the theory is concealed in your data for you to discover. Coding makes visible some of its components. Memoing adds the relationships which link the categories to each other.* »[209]

[208] https://sourceforge.net/projects/nodepad/
[209] www.scu.edu.au/schools/gcm/ar/arp/grounded.html

NooJ[210]

1. Caractéristiques générales

Nooj est un logiciel libre créé par Max Silberztein concepteur d'Intex. Il est présenté comme le successeur de ce dernier.

Il est disponible pour de nombreuses langues : romanes, germaniques, slaves, sémitiques, asiatiques, etc.

Il peut traiter des corpus de plusieurs centaines de textes.

2. Domaines d'application et fonctionnalités

NooJ est un environnement de développement linguistique « *permettant de construire, de tester et de gérer des descriptions formalisées à large couverture des langues naturelles, sous forme de dictionnaires et de grammaires électroniques, ainsi que de développer des applications du TAL* ». [211]

Les dictionnaires et grammaires sont appliqués au corpus dans le but d'identifier des motifs morphologiques, lexicaux, syntaxiques et de marquer les mots simples et composés.

NooJ contient notamment[212] :

- des mini-applications pédagogiques ;
- des outils permettant la consultation et l'extraction d'informations à partir du corpus "brut", ne comportant pas de balises spécifiques ;
- un analyseur morphologique qui permet d'effectuer des recherches et des traitements dans les textes à partir d'expressions régulières intégrant des formes, des lemmes, des catégories syntaxiques ou toute information lexicale ;
- des outils permettant d'appliquer des grammaires (représentées graphiquement) à un corpus ;
- des outils permettant de construire et de gérer des concordances.

R[213]

1. Caractéristiques générales

R est un logiciel libre.

Il constitue une implémentation du langage S développé aux laboratoires Bell.

Il est possible de l'utiliser sous Unix, Windows et MacOS.

---

[210] http://www.nooj4nlp.net/ ; M. Silberztein et A. Tutin, « NooJ, un outil TAL pour l'enseignement des langues. Application pour l'étude de la morphologie lexicale en FLE », http://alsic.u-strasbg.fr/v08/silberztein/alsic_v08_20-rec11.htm

[211] M. Silberztein et A. Tutin, « NooJ, un outil TAL pour l'enseignement des langues. Application pour l'étude de la morphologie lexicale en FLE », http://alsic.u-strasbg.fr/v08/silberztein/alsic_v08_20-rec11.htm

[212] http://www.nooj4nlp.net/

[213] http://www.r-project.org/

2. Domaines d'application et fonctionnalités

R est une suite intégrée de logiciels permettant la manipulation de données, des calculs et un affichage graphique des résultats.

R comprend un logiciel statistique « permettant d'appliquer des modèles de distributions statistiques à des "nuages de points" (ou plus simplement des tableaux de fréquences) ».[214]

Il possède de plus une librairie dédiée à l'analyse des données textuelles.[215] Celle-ci offre des fonctions de segmentation du corpus, de réduction des formes accentuées, de lemmatisation, de création du tableau lexical, d'analyse (fréquences, probabilités).[216]

SATO (Système d'Analyse de Textes par Ordinateur)[217]

1. Caractéristiques générales

Le logiciel SATO a été développé par le Service d'Analyse de Textes par Ordinateur de l'Université du Québec à Montréal (SATO).

Il est disponible pour des plates-formes informatiques de type IBM-PC et compatibles.

Il est également disponible en ligne après inscription.

2. Domaines d'application et fonctionnalités

SATO permet différents types d'analyse : l'annotation de documents multilingues, le repérage sur mesure des éléments du texte et l'analyse qualitative ou quantitative du document ou de ses parties.

Il offre notamment les fonctionnalités suivantes[218] :

- un langage de requête assurant le repérage systématique de segments textuels définis par l'utilisateur au moment de la requête ;
- le repérage de contextes par des patrons de concordance ;
- la production d'index ;
- la constitution d'inventaires lexicaux triés alphabétiquement ou numériquement, ou selon tout autre système de description ;
- la catégorisation de mots, de mots-composés ou de locutions ;
- la définition de variables pour effectuer des dénombrements multiples et analyses lexicométriques ;
- des fonctions pour constituer et mettre à jour des dictionnaires avec, si nécessaire, des dispositifs pour la dérivation morphologique ;
- l'analyse des cooccurrences ;
- la gestion des formats d'affichage ;
- un indice de lisibilité ;
- un mode assisté de mise au point de scénarios automatiques.

Sémato[219]

[214] http://analyses.ishs.ulg.ac.be/logiciels/opencaqdas.html
[215] http://wwwpeople.unil.ch/jean-pierre.mueller/
[216] J.-P. Müller, « La librairie ttda: tools for textual data analysis », 5 mars 2004, http://wwwpeople.unil.ch/jean-pierre.mueller/ttda-fr.pdf
[217] http://analyses.ishs.ulg.ac.be/logiciels/sato.html ; http://www.ling.uqam.ca/sato/ ; http://www.ling.uqam.ca/ato/index.html
[218] http://www.ling.uqam.ca/sato/

1. Caractéristiques générales

Sémato est un logiciel d'analyse sémantique de documents textuels développé par Pierre Plante, Lucie Dumas et André Plante, à l'Université du Québec à Montréal comme SATO.

Il est disponible en français ou en anglais.

Sémato est un logiciel d'analyse en ligne.

2. Domaines d'application et fonctionnalités

Ses domaines d'application sont multiples : analyse de focus groups, de questions ouvertes dans les sondages, d'entrevues dirigées, semi-dirigées ou libres, de corpus littéraires ou socio-politiques, d'articles de journaux, etc.

Sémato permet l'identification automatique des éléments saillants du contenu textuel. [220]

« *Sémato* offre plusieurs outils qui permettent d'explorer, de repérer et d'analyser le contenu du corpus. Son utilisation s'articule principalement autour :

- de l'élaboration des thèmes qui sont utilisés pour le repérage et l'analyse d'extraits du corpus.
  Les thèmes correspondent à un regroupement d'indicateurs et sont, en quelque sorte, l'équivalent des codes et des catégories que l'on utilise habituellement en analyse qualitative. Il s'agit donc d'un élément-clé des analyses assistées par *Sémato*.
- d'outils de repérage et d'analyse en direct qui permettent l'exploration du corpus. » [221]

Sur base de cette analyse thématique, le logiciel autorise de multiples analyses croisées entre les variables externes qui caractérisent les éléments textuels (auteur, genre, date, domaine, etc.) et les éléments du contenu trouvés de façon automatique ou assistée.

Il offre un outil d'analyse des cooccurrences.

TamsAnalyser (TAMS = Text Analysis Mark-up System)[222]

1. Caractéristiques générales

TamsAnalyser est un logiciel libre développé par Matthew Weinstein.

Il est disponible sous Mac Os X et Linux.

2. Domaines d'application et fonctionnalités

Ce logiciel peut être utilisé dans des domaines variés : études de médias, anthropologie, éducation, sociologie.

---

[219] http://fable.ato.uqam.ca/guidexpert-ato/gea.asp
[220] J. Saint-Charles, P. Mongeau, « Guide d'initiation », Département de communication sociale et publique, UQAM, Septembre 2006, http://www.er.uqam.ca/nobel/r32700/Semato/Guide_Semato.pdf
[221] J. Saint-Charles, P. Mongeau, « Guide d'initiation », Département de communication sociale et publique, UQAM, Septembre 2006, http://www.er.uqam.ca/nobel/r32700/Semato/Guide_Semato.pdf
[222] http://tamsys.sourceforge.net/

TamsAnalyser permet l'analyse des thématiques textuelles.
TAMS (Text Analysis Markup System) est une convention d'identification de thèmes dans des textes. Celle-ci a été conçue pour la recherche ethnographique et sur le discours.

Il s'agit d'un outil permettant le codage (ethnographique, à partir d'une liste fournie) et ensuite la manipulation (extraction, analyse, sauvegarde des informations codées) de segments de texte.

Il peut toutefois être utilisé pour l'analyse de discours en sciences sociales et culturelles

Tetral*o*gie[223]

1. Caractéristiques générales

Tetralogie a été développé par l'Institut de Recherche en Informatique de Toulouse.
Il est disponible sous environnement UNIX et est accessible à partir de terminaux X, PC, Macintosh équipés d'un émulateur X.

Le logiciel TETRALOGIE est un des éléments essentiels de la station bibliométrique "ATLAS".

2. Domaines d'application et fonctionnalités

Ses domaines d'applications sont la veille technologique, la veille stratégique, le data mining, l'analyse exploratoire des données.

TETRALOGIE est un logiciel d'études bibliométriques, dont l'objectif est de découvrir les grands axes d'intérêt d'un corpus et d'accéder à l'information endogène. Il est basé sur l'analyse exploratoire des données et les méthodes de classification automatique.

Il intègre des méthodes suivantes : l'analyse en composantes principales, l'analyse factorielle des correspondances, la classification ascendante hiérarchique, la classification par partitions, ainsi que des méthodes d'analyse de l'évolution (relative et absolue) et d'analyse relationnelle des données portant sur les liens (secondaires, ternaires...).

TETRALOGIE va sélectionner les données dans le corpus par le biais de séparateurs insérés de façon semi-automatique.

Plusieurs procédures statistiques sont proposées : élaboration de filtres et de thésaurus mono ou multitermes, élaboration de dictionnaires de synonymes par spécialité, de dictionnaires de mots vides, application de techniques de croisement des informations pour obtenir des matrices de fréquence, de présence-absence ou de cooccurrences (simples et multiples) sur lesquelles porteront ensuite les analyses.

[223] http://mist.univ-paris1.fr/logiciel/tableau/tetratab.htm ; http://atlas.irit.fr/TETRALOGIE/tetrajeu.htm

Transcriber[224]

1. Caractéristiques générales

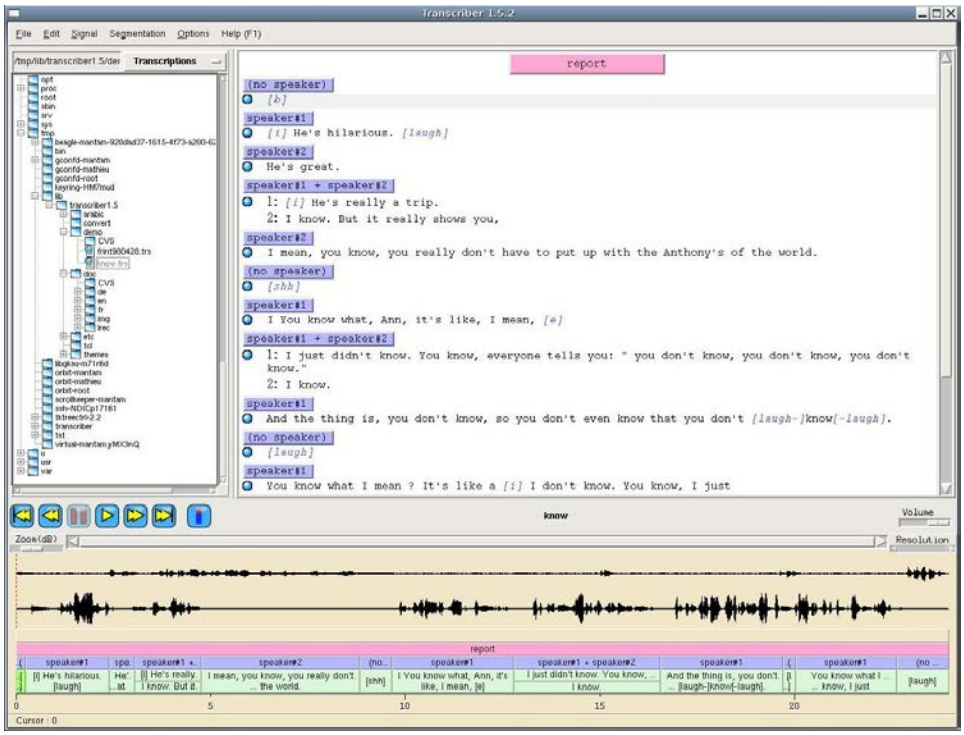
Transcriber est un logiciel libre développé par Mathieu Manta, Fabien Antoine, Sylvain Galliano et Claude Barras.

Il a été testé sous Linux, MacOS X et Windows XP.

2. Domaines d'application et fonctionnalités

Transcriber est d'un outil d'aide à l'annotation d'entretien ou de focus group non retranscrits.
Il permet la segmentation de longs entretiens enregistrés et leur transcription, ainsi que la « labellisation » des tournants dans l'entretien, des changements de sujets et des conditions acoustiques.

Exemple d'écran :



Tri-Deux[225]

1. Caractéristiques générales

Tri-Deux est un logiciel libre développé par Philippe Cibois (Université d'Amiens).

Il fonctionne sous Windows (98 à XP).

---

[224] http://trans.sourceforge.net/en/presentation.php

[225] Jacques Jenny, « Méthodes et pratiques formalisées d'analyse de contenu et de discours dans la recherche sociologique française contemporaine. Etat des lieux et essai de classification. », Article publié dans le *Bulletin de Méthodologie Sociologique (B.M.S.),* n° 54, mars 1997, p.64-112, http://pageperso.aol.fr/jacquesjenny/ATBMS.htm ; http://perso.orange.fr/cibois/SitePhCibois.htm

2. Domaines d'application et fonctionnalités

Tri-Deux est un logiciel de dépouillement d'enquête.

Différentes techniques sont intégrées : des techniques simples comme les tris croisés ou plus complexes comme l'analyse factorielle, des méthodes post-factorielles ou la régression sur données d'enquête.

Tropes[226]

1. Caractéristiques générales

Tropes est un logiciel d'analyse de textes développé par la société Acetic[227], firme française spécialisée dans les logiciels d'analyse et de traitement de l'information.

Il existe en versions française et anglaise.

Le logiciel est disponible à partir de Windows 98.

Il est complété par les logiciels Zoom et Index.
Zoom est un « moteur de recherche sémantique et d'analyse documentaire, fondé sur la compréhension du contenu qu'il a à traiter ».[228]
Index est un « composant logiciel permettant d'extraire, de formater, de classer et de réutiliser l'information recueillie au cours d'un traitement linguistique poussé (analyse sémantique) d'une base documentaire (une collection de fichiers textes stockés sur un poste de travail ou un serveur) qui peut comprendre des millions de documents ».[229]

2. Domaines d'application et fonctionnalités

Les domaines d'application de Tropes sont variés : sondages, enquêtes, veille concurrentielle, veille sociétale, intelligence économique, études comportementales, études sociologiques, analyse de discours politiques.

Tropes est un outil d'analyse sémantique de textes. Pour effectuer cette analyse, il effectue un traitement visant à affecter tous les mots significatifs dans des catégories, à analyser leur répartition en sous-catégories (catégories de mots, classes d'équivalents), à étudier leur ordre d'arrivée à la fois à l'intérieur des propositions (relations, actants et actés), et sur l'intégralité du texte (graphe de répartition, rafales, épisodes, propositions remarquables).[230]

Il offre un outil permettant de résoudre les ambiguïtés lexicales et sémantiques et est basé sur des méthodes d'analyse telles l'analyse propositionnelle du discours, l'analyse cognitivo-discursive et la méthode des rafales.

Tropes est architecturé autour d'un moteur d'analyse linguistique et s'appuie sur des dictionnaires et des réseaux sémantiques.

---

[226] http://www.acetic.fr/*Tropes*.htm ; http://mist.univ-paris1.fr/logiciel/tableau/troptab.htm
[227] http://www.acetic.fr/index.htm
[228] http://www.acetic.fr/zoom.htm
[229] http://www.acetic.fr/indexation-semantique.htm
[230] http://www.acetic.fr/*Tropes*.htm

L'analyse par Tropes s'effectue en six étapes :

- découpage des phrases et des propositions,
- levée d'ambiguïté des mots du texte,
- identification des classes d'équivalents sémantiques,
- statistiques, détection des rafales et des épisodes[231],
- détection des propositions remarquables (contraction du texte),
- mise en forme et affichage du résultat.

Unitex[232]

1. Caractéristiques générales

Unitex, logiciel libre, a été développé sous la direction du linguiste Maurice Gross.

Il est capable de traiter n'importe quelle langue codifiée en Unicode.

Il a été testé sous Windows, Linux, Mac OS X.

2. Domaines d'application et fonctionnalités

Unitex permet d'étudier, dans un corpus, les concordances d'expressions (co-référence) et de travailler sur l'ambiguïté.

Il travaille sur base de dictionnaires et de grammaires intégrées.

Il offre des fonctions comparables à celles d'Intex.

Weblex[233]

1. Caractéristiques générales

Weblex est un logiciel de lexicométrie directement accessible par Internet.

2. Domaines d'application et fonctionnalités

Weblex offre[234] :

- « l'étiquetage automatique du texte par des propriétés linguistiques et sémantiques : morphosyntaxe, lemmatisation, ... ;
- un moteur de recherche très complet qui permet une analyse *locale* et fine du contexte d'apparition de chaque mot ;

[231] Une rafale regroupe des occurrences de mots (contenus dans une classe d'équivalents sémantiques) ayant tendance à arriver avec une concentration remarquable dans une partie limitée du texte (en début, milieu ou fin du texte mais jamais sur son intégralité).
Un épisode correspond à une partie du texte où un certain nombre de rafales se sont formées. Ils constituent des blocs d'argumentation, représentatifs de la structure du discours observé. http://www.acetic.fr/*Tropes*.htm

[232] http://www-igm.univ-mlv.fr/%7Eunitex/index.html ; http://www-igm.univ-mlv.fr/~unitex/ ; http://www-igm.univ-mlv.fr/~paumier/DEA/Cours%205%20-%20Introduction%20a%20Unitex.pdf

[233] http://weblex.ens-lsh.fr/wlx/

[234] http://weblex.ens-lsh.fr/wlx/

- des outils de mesure statistique contrastifs ou non qui offrent différentes synthèses rapides et *globales* de l'usage de son vocabulaire (cooccurrences, spécificités, AFC, ...) et de divers évènements textuels. »

Il offre également la recherche de cooccurrences.

Wordstat[235]

1. Caractéristiques générales

Wordstat est commercialisé par la société Provalis Research.

Ce logiciel est disponible en anglais, français, italien et espagnol.

2. Domaines d'application et fonctionnalités

Wordstat permet l'analyse textuelle de documents variés : réponses à des questions ouvertes, interviews, titres, articles de journaux, discours, communications électroniques, etc.

Il peut être utilisé via des dictionnaires de catégorisation existant ou en mode manuel. Il peut aussi participer à la validation de nouveaux dictionnaires

Il inclut différents outils d'analyse de donnés qui permettent d'explorer les relations entre le contenu du corpus étudié et l'information contenue dans des variables de catégorisation ou numériques telles l'âge, le genre, l'année de publication, etc.

Ces relations sont identifiées via un clustering hiérarchique et une analyse à échelle multidimensionnelle.

Les fonctionnalités suivantes sont notamment offertes[236] :

- Catégorisation des mots ou des phrases sur base de dictionnaires existants ou créés par l'utilisateur ;
- Analyse fréquentielle des mots-clés, phrases, catégories dérivées ou concepts ;
- Analyse des cooccurrences ;
- Fonction de repérage des mots-clés ;
- Analyse des similarités entre cas ou documents ;
- Comparaisons ;
- Classification automatique des textes ;
- Analyse du contexte de mots-clés ;
- Statistiques.

[235] http://www.provalisresearch.com/wordstat/wordstat.html
[236] http://www.provalisresearch.com/wordstat/wordstat.html

4. Texte analysé

Le "Web 2.0", nouvelle bulle Internet ?

Mathieu Van Overstraeten

Mis en ligne le 27/02/2007

Benoît Lips, ex- "gourou de l'Internet", dirige l'agence web DAD, devenue LBI.
Ayant survécu à la première bulle Internet, il craint une nouvelle euphorie excessive.

Entretien

Il y a une dizaine d'années, Benoît Lips se rendit célèbre en expliquant le b.a.-ba du web dans un best-seller intitulé "Internet en Belgique". Considéré à l'époque comme un des premiers "gourous" de ce nouveau média, il signa alors des chroniques hebdomadaires dans "La Libre Entreprise" et fut régulièrement interrogé à la radio et à la télévision. Un succès qui n'empêcha pas cet ingénieur civil de formation de rester les deux pieds sur terre, même au plus fort de l'euphorie Internet. La preuve : sa société, DAD (Digital age design), fondée en 1995 et spécialisée dans la conception de sites web, est toujours bel et bien là. Rachetée par Belgacom en 1997, puis par le groupe suédois LBI International il y a 18 mois, elle vient d'être rebaptisée LBI afin de mieux s'intégrer dans le réseau européen de sa maison mère. Employant quelque 80 personnes, elle compte engager une dizaine de collaborateurs supplémentaires en 2007 et surtout franchir la barre symbolique des 10 millions d'euros de chiffre d'affaires.

Pourquoi votre société a-t-elle changé de mains, passant du groupe Belgacom au groupe LBI ?

Pour deux raisons. D'une part, en faisant son entrée en Bourse, Belgacom a logiquement cherché à rationaliser ses participations. D'autre part, même si nous avions une grande marge de manœuvre au sein du groupe Belgacom, il y avait un manque de compréhension et de synergie entre nos métiers. Certes, Belgacom nous a permis d'avoir une vision à long terme à un moment où, comme d'autres, nous aurions pu être trop fougueux et jeter notre argent par les fenêtres. Mais en même temps, au sein de Belgacom, nous étions tout petits, alors que chez LBI, non seulement nous représentons à peu près 10 pc des effectifs et du chiffre d'affaires mais, en plus, nous sommes actifs dans un métier comparable. Ce qui est enrichissant, car la Belgique n'est pas toujours très en avance au niveau Internet. Dans le groupe LBI, la filiale italienne, par exemple, est très avancée dans la TV interactive, alors que le bureau de Berlin est très fort dans le "branding". C'est vivifiant de pouvoir voir ce qui se passe dans le reste de l'Europe. Je n'ai d'ailleurs jamais autant voyagé qu'aujourd'hui.

Vous vous définissez comme le leader belge du marketing digital. Est-ce que cela signifie que vous concurrencez aussi les agences de pub traditionnelles ?

Il est clair que l'Internet fait aujourd'hui définitivement partie de la stratégie marketing et communication des entreprises, et que le monde de la pub ne sait pas encore très bien comment répondre à ce mouvement, si ce n'est par la création de quelques micro-cellules Internet. Cela dit, il y a une réelle prise de conscience des grands groupes de communication par rapport au marketing digital, et je m'attends à des acquisitions dans ce domaine à l'avenir.

Comment une société comme la vôtre réagit-elle à l'émergence du "Web 2.0", dans lequel le contenu Internet est de plus en plus souvent généré par les utilisateurs, à l'image de sites tels que YouTube ou MySpace ?

L'engouement actuel me fait un peu peur, d'autant plus qu'on est à nouveau dans une période d'envolée boursière de certaines sociétés Internet. Il y a, autour du "Web 2.0", un côté "hype" et une surexposition qui me paraissent suspects. Les sociétés pensent qu'il suffit de poster un petit clip sur YouTube pour que des internautes du monde entier aillent le voir, tout comme elles croyaient il y a dix ans qu'il suffisait d'avoir un site Internet pour que tout le monde le consulte.

A vous entendre, le "Web 2.0" ne serait donc pas la révolution que certains annoncent...

Le concept me séduit, mais il ne me paraît pas si révolutionnaire que cela, dans la mesure où l'Internet est collaboratif par essence. Le "Web 2.0" rend simplement les interactions plus jolies, plus design et plus simples. Cela dit, je suis d'accord pour dire que le fait de laisser aux internautes la possibilité de s'exprimer peut être un outil marketing très puissant. Moi-même, je vais toujours lire les commentaires des autres internautes avant de choisir un resto ou un lieu de voyage par exemple. Mais encore faut-il être capable d'utiliser ces nouveaux outils à bon escient. Je pense donc qu'il reste un grand travail de sensibilisation à faire pour que les entreprises comprennent exactement ce que signifient ces nouvelles tendances.

Cet article provient de **http://www.lalibre.be**

5. Le calcul des spécificités[237]

Cette méthode statistique est due à P.Lafon qui, dans les années 70, a proposé d'appliquer la distribution hypergéométrique à la question de la répartition des "formes" dans un corpus.

« Cette méthode permet de mesurer les variations de la fréquence dans un corpus découpé en parties et, en fonction d'un seuil choisi par l'analyste, il indique si la fréquence observée dans telle ou telle partie peut-être considérée comme normale ou non. Dans ce dernier cas, P. Lafon propose de baptiser cette forme "spécifique" (de la partie considérée). »

L'application de la formule suppose les notations suivantes :

- T : la longueur du corpus (nombre total de mots de celui-ci)
- ti : la longueur de la partie i
- f : la fréquence absolue d'une forme dans le corpus entier.
- fi : fréquence absolue d'une forme dans la partie i
- X : la variable aléatoire mesurant le nombre d'apparitions d'une forme dans la partie considérée.

On calcule une probabilité pour qu'une forme de fréquence f apparaisse k fois dans la partie i :

$$(1) \quad P(X=k)=\frac{\binom{f}{k}\binom{T-f}{t_i-k}}{\binom{T}{t_i}}$$

Cette probabilité atteint son maximum à l'espérance mathématique, c'est-à-dire la "fréquence attendue dans la partie i", qui doit être nécessairement un entier tel que :

$$\frac{(f+1)(t_i+1)}{T+2} - 1 \leq f'_i \leq \frac{(f+1)(t_i+1)}{T+2}$$

Si la fréquence observée n'est pas égale à cette fréquence attendue, on peut se demander si l'écart entre les deux valeurs est ou non significatif.

La probabilité pour qu'on rencontre une fréquence telle que celle observée, sera :

premièrement, avec fi>f'i :

$$S^{+} = P(X \geq f_i) = \sum_{k=f_i}^{Min(f;t_i)} P(X=k)$$

Si S+ est plus petit qu'un seuil choisi par l'opérateur — généralement 0,05 ou 0,01 —, on parle alors de *spécificité positive* : le mot est, d'après le calcul hypergéométrique, significativement "sur-employé" dans la partie considérée.

deuxièmement, avec fi<f'i :

[237] C. Labbé et D. Labbé, « Que mesure la spécificité du vocabulaire ? »

$$S^- = P(X \leq f_i) = \sum_{k=0}^{fi} P(X=k)$$

Dans le cas où S- est plus petit que le seuil choisi, on parle de *spécificité négative* : le mot est significativement "sous-employé" dans la partie considérée.

En résumé, le calcul de la spécificité permet de déterminer si un terme est sur- ou sous-représenté dans les différentes parties du texte.

6. Interviews

6.1. Interview de Michaël Petit – 23 janvier 2007

1. **Quel est le type de données dont vous vous servez ?**
2. **Quelles sont vos pratiques en matière d'analyse de texte ?**

Il serait utile d'avoir des logiciels permettant de contrôler le plagiat dans les travaux d'étudiants (plagiat entre groupes d'étudiants ou de travaux existants). On ne le fait pas actuellement.

Il serait aussi utile d'avoir des logiciels permettant de faire des recherches sur un sujet particulier (dans le cadre de la recherche ou de l'enseignement). Actuellement, on utilise des mots-clés sur Google. Beaucoup de documents sont trouvés, mais ce ne sont pas toujours les plus adaptés. On se situerait dans une démarche quantitative.
Il faut souvent procéder par tâtonnement en cherchant également par des synonymes ou des mots connexes du thème que l'on veut explorer.

Des logiciels d'indexation des documents du disque dur pourraient aussi avoir leur utilité. On est à nouveau dans une perspective quantitative, mais éventuellement étendue à une base qualitative / syntaxique.
On est à nouveau confronté au problème des mots connexes.
Ces logiciels devraient permettre de retrouver un document attaché à un mail ou rangés à un endroit auquel on ne pense plus.
Mais il paraît impossible de laisser le logiciel classer lui-même les documents. L'organisation ne peut rester que manuelle via l'arborescence des fichiers.

On a parfois des interviews dans le cadre de contrats de recherche. Mais on n'utilise pas d'outil pour les analyser.
On a surtout des cahiers des charges pour les projets de recherche.

En ce qui concerne les interviews, elles servent pour la création de modèles. On les analyse manuellement pour en extraire des informations qui peuvent être introduites dans des modèles.
Ex. : - modèle UML : diagrammes de classe et diagrammes d'activité
- modèle d'objectif : missions des services d'urgence – objectifs en terme de qualité

On a aussi d'autres modèles basés sur l'observation d'acteurs.
Ex. : on suit des médecins ou des infirmiers et on note tout ce qu'ils font dans un langage structuré, dans des grilles séquentielles et événementielles.
On peut alors créer des diagrammes de séquence et delà, des diagrammes d'activité.
On extrait des connaissances pour les traduire en modèle UML.
On observe des scénarios : des séquences de choses. Et on essaie de généraliser à partir de plusieurs.
Ce n'est pas thématique, c'est plutôt séquentiel, une série d'étapes. Il y a les acteurs, les activités observées et les données manipulées.
Eventuellement, si on a une série de scénarios au format textuel, on peut chercher des points communs entre eux via un logiciel.
A partir de documents en langage naturel, il s'agit d'extraire des éléments pour en faire un modèle structuré. Il faudrait des logiciels qui puissent faire cela.

On a aussi des textes très structurés : les cahiers des charges.
C'est nous-même qui les faisons, pas le client.
Au niveau de la structure, on utilise des templates (Volere, STD 830).
A chaque section prédéfinie correspond un certain type de contenu, ex. : le lien entre le système à développer et l'environnement fonctionnel et non fonctionnel.
En ce qui concerne les logiciels d'analyse, on pourrait vérifier que le contenu d'une section est adapté à ce qui est demandé via le vocabulaire utilisé.
Mais c'est quand même moins intéressant car on a des modèles très structurés.

Des logiciels qui pourraient fournir des schémas entité-association pourraient être utiles, ou décrire les fonctions attendues d'un système en analysant les exigences mentionnées dans des interviews.

En ce qui concerne la recherche (domaine business par exemple), on applique des modèles pour structurer le discours. On a des langages spécifiques.
On est face à un langage informel, peu structuré, mais certaines choses reviennent souvent. L'analyse doit permettre de détecter des concepts que l'on a dans le modèle. Cela facilite la transition vers un langage plus formel.
Il serait intéressant de voir comment des outils pourraient être utiles pour cela.
On aurait besoin d'un outil d'analyse dirigé par un vocabulaire déjà existant.
On se base sur une ontologie existante : un ensemble de concepts hiérarchisés et ayant des relations entre eux. Il s'agit d'un méta-modèle.

Des logiciels seraient aussi utiles pour créer une description d'un domaine de recherche, une ontologie reprenant tous les concepts d'un domaine et leurs liens.
On peut citer le projet Interop sur l'interopérabilité, dans lequel on essaie de définir le domaine de recherche.
Ex. : au niveau des données, il faut un format compatible ; au niveau du code, il y a échange de messages.
On peut procéder de deux manières pour faire cette ontologie.
On peut réfléchir dessus et la définir manuellement.
Ou on peut analyser les textes existants sur le domaine. On prend l'ensemble des documents des spécialistes du domaine et on procède à une extraction des termes pertinents via un logiciel d'analyse.
Il s'agit à la fois d'une démarche qualitative et d'analyse des relations entre les concepts.
L'idée de contraste entre vocabulaires est aussi importante : un terme n'est pertinent pour un domaine que s'il est présent dans les articles du domaine et absent dans les articles d'autres domaines.
L'extraction de termes se baserait sur des mesures classiques. De là, on aurait un glossaire du domaine (un lexique plat). À partir de là, il faudrait un outil permettant de retrouver les définitions des termes dans les articles ou ailleurs.
Cette liste avec les définitions serait validée par des experts.
Ensuite, ces termes seraient placés dans une taxonomie[238] (du général vers le particulier), par inclusion des termes particuliers dans un terme plus général.
Cette taxonomie serait revalidée, et on obtiendrait le glossaire du domaine.

[238] Une taxonomie est une classification sous forme d'arbre sans cycle et linéaire, on descend dans l'arbre du plus général au plus spécial.
Dans une ontologie, il y a des liens entre les concepts en plus.

Toujours dans le cadre de la recherche, si on a un domaine nouveau avec un petit groupe et un thème limité, il est intéressant de voir comment on peut structurer les connaissances que l'on a déjà grâce à un outil. Il s'agirait de créer un modèle général pour le domaine.
L'analyse de texte devrait permettre d'extraire tous les concepts importants du domaine sur base d'articles de personnes de l'équipe ou d'autres articles.
Le logiciel devrait pouvoir extraire les termes et leurs relations afin de faire l'ontologie du domaine.
Un deuxième outil devrait permettre de découvrir des informations liées à l'ontologie.
Ex. : analyser des sites web pour voir s'ils sont en rapport avec l'ontologie et s'ils sont pertinents.
De plus, il devrait être possible d'enrichir l'ontologie à partir des nouvelles découvertes.
Le logiciel devrait pouvoir analyser la pertinence du document et l'intégrer dans l'ontologie.
Cela devrait pouvoir s'appliquer à toutes sortes de documents, les e-mails par exemple.
Vu le grand nombre de newsletters et de mailing-list, il est difficile de lire tout.
Un logiciel d'analyse devrait pouvoir analyser ces mails afin de déterminer s'ils sont intéressants pour le chercheur.
Il devrait pouvoir faire le matching entre l'ontologie et l'information textuelle brute. Et si l'information est pertinente, il devrait pouvoir l'extraire pour l'intégrer à ce qui existe.
Actuellement, tout cela se fait manuellement.

Mais tout ne pourrait pas être automatique car un mot-clef peut avoir plusieurs significations et il peut y avoir plusieurs mots pour un concept. Il faudrait des systèmes intelligents avec détection des synonymes et des relations entre concepts.

3. **Y a-t-il des contraintes en ce qui concerne la taille des données ?**

La taille des données est variable :
- dans le cadre de la création d'ontologie, on est face à de gros volumes de données ; on peut prendre tous les articles de conférences pendant quelques années ; cela fait une centaine d'articles ;
- en ce qui concerne l'organisation des connaissances, c'est moins important ; on prend les articles des personnes du groupe ;
- les cahiers des charges et les textes stratégiques pour les entreprises sont plus petits ;
- les documents scientifiques également.

4. **Entrées / sorties ?**
5. **Quel type d'affichage des résultats souhaitez-vous ?**

Des formats d'entrée différents sont intéressants : PDF, Word, html, txt, PS, Latech éventuellement.

Au niveau des sorties, il devrait être possible de visualiser graphiquement l'ontologie résultante et de la valider.

En ce qui concerne l'interface, on devrait pouvoir fournir des documents au système et choisir des options d'analyse (ex. : analyser le texte en utilisant telle ontologie).

Au niveau des résultats, on devrait voir l'ontologie sous forme de graphe avec une mesure (sous forme de degré) de la pertinence des termes du document par rapport à l'ontologie.

Il devrait être possible de sortir l'ontologie sous format OWL.

A partir du document brut en langage naturel, il devrait être possible de sortir un modèle dans un format conforme.

On a au départ un modèle avec des concepts reliés. L'analyse textuelle devrait permettre la découverte de mots-clés qui seraient des instanciations des concepts du modèle. Le logiciel devrait afficher les liens entre mots-clés et concepts.

6. **Le système d'exploitation a-t-il une importance ?**

Non. De même, peu importe s'il s'agit de logiciels à installer ou en ligne.

7. **Quelle est la langue principale ?**

Dans la recherche, 99% des documents sont en anglais. Les cahiers des charges sont en français et parfois en anglais.
Ce serait bien d'avoir un système multilingue.

Au niveau du logiciel, il devrait être en français ou en anglais.
L'anglais serait plus cohérent vu que les données sont majoritairement en anglais et que le logiciel pourrait être utilisé par des non francophones plus facilement.

6.2. Interview de Patrick Heymans – 23 janvier 2007

1. **Quel est le type de données dont vous vous servez ?**
2. **Quelles sont vos pratiques en matière d'analyse de texte ?**

* Un projet où l'on pourrait avoir besoin de logiciels d'analyse de texte est celui de l'analyse de risque de sécurité pour les SI.
L'état de l'art est très riche mais divergent.
D'un côté, on a les standards d'analyse des risques venant des organisations de standardisation (surtout des industriels). Ces documents ont un lexique avec une terminologie standard. Il s'agit de documents méthodologiques avec une analyse des risques et des contre-mesures pour éviter ou partager les risques. On identifie les risques managériaux et opérationnels du système quand il tourne.
Il s'agit de documents méthodologiques verbeux.
De l'autre, au niveau de l'ingénierie des exigences des SI, on a des langages et des extensions de langages existants permettant de gérer les risques.
Ex. : au niveau des use cases, on a l'extension des misuse cases pour les scénarios que l'on veut éviter.
Le problème est que la terminologie utilisée est différente selon le langage et les standards.

Le travail consiste à essayer d'uniformiser la terminologie. On regarde tous les standards et on essaie de faire une table d'alignement des concepts. Il s'agit de voir les différences de vocabulaire pour un même concept.

On procède en lisant tous les standards et en essayant de retrouver les définitions des termes. Il y a parfois un lexique, parfois il faut se baser sur des définitions implicites.

On fait un matching sémantique.
Tout se fait manuellement.

Un problème vient de l'ambiguïté des définitions. Est-ce qu'il s'agit bien de la même chose dans les deux standards ? S'il y a des différences, de quelle nature sont-elles ?
On se réfère au bon sens pour répondre à ces questions.

Le but est d'exprimer le langage naturel avec certitude.

Après avoir identifié ces concepts et leurs définitions, on peut construire un méta-modèle représentant de manière plus formelle les concepts d'analyse de risque et leurs liens.
Celui-ci est validé par des experts.

On a une représentation sous forme de diagramme de classe.
On peut procéder à des comparaisons avec les méta-modèles des langages servant à définir les modèles de sécurité.
Ex. : le langage KAOS (méta-modèle) – on cherche des liens entre les concepts de KAOS et le méta-modèle d'analyse de risque.
Toutes ces comparaisons se font à la main.

L'évaluation vise à voir à quel point tel langage couvre les concepts de l'analyse de risque. Est-ce que tel langage orienté sécurité couvre les concepts nécessaires à l'analyse de risque ?

L'idée sous-jacente est que les personnes qui font de la sécurité et celles qui font de l'analyse de risque ne se parlent pas. Il n'y a donc pas de garantie que les exigences permettent d'éliminer ce qui est considéré comme un risque important pour une société.

Cela prend beaucoup de temps, donc un peu d'automatisation serait utile.

* Un autre projet : UEML (Interop).
Il vise à définir la sémantique ontologique des langages de modélisation.

L'ontologie est un concept issu de la philosophie et notamment du philosophe Mario Bunge qui a voulu identifier les concepts de base pour décrire et classifier le monde physique. Sa théorie se base sur la notion de « chose ».
L'idée est de reprendre cela pour modéliser les concepts des SI.
Cela permet d'évaluer les langages de modélisation. Son-ils assez riches pour exprimer tout ce qu'on veut ?

Au niveau informatique, le projet UEML est une utilisation d'une extension de l'ontologie aux SI. On utilise les modèles Bunge-Wand-Weber.
On prend la classification et on en fait un modèle conceptuel, on prend le méta-modèle d'un langage qu'on veut étudier, et on cherche des liens entre les deux.
A la fin, on obtient la sémantique ontologique des concepts. Pour chaque concept du langage, on a un concept de Bunge-Wand-Weber.

On voudrait pouvoir automatiser l'analyse des correspondances. Est-ce que deux langages représentent la même chose ? Qu'est-ce qui correspond au langage A dans le langage B ?

On rencontre différents problèmes :

- faire le méta-modèle d'un langage est très complexe ;
- tracer des liens est très arbitraire les concepts peuvent être compris différemment.

Pour y remédier, on a distribué 3-4 langages dans différents laboratoires et on a demandé de faire l'évaluation du mapping fait par d'autres afin d'obtenir un consensus.

Il y a aussi le problème des ambiguïtés, mais qui ne peut pas être résolu par le méta-modèle.

* Travail de mémorants : outil de documentation des exigences pour faire des cahiers des charges - GenSpec
Il y a des problèmes d'ambiguïtés : utilisation de termes dans des sens différents.
Donc, on ajoute des dictionnaires au cahier.
Un outil utile devrait pouvoir repérer les termes ambigus. Et s'il y a des synonymes, il devrait pouvoir le signaler.
Cela pousserait à utiliser une terminologie uniforme.
Le logiciel devrait souligner les termes du lexique et renvoyer aux définitions.

* En ce qui concerne des interviews, on n'en traite pas directement. C'est une des techniques en ingénierie des exigences, mais on n'analyse pas directement les retranscriptions.

* En ingénierie des exigences, on a des documents informels dont on veut faire un document conceptuel. Il s'agit d'extraire des modèles conceptuels.
On a fait des tentatives, mais sans succès jusqu'à présent.
Le problème est de comprendre la sémantique de ce qu'il y a dans le texte. C'est très précis, il faut aller dans les détails.
Il faudrait des ordinateurs qui puissent réfléchir comme nous.

* Il y a aussi l'utilisation de langages naturels contrôlés avec l'imposition d'une syntaxe restreinte pour les descriptions. A partir de là, on peut tirer des modèles conceptuels, et des diagrammes. Mais c'est très strict.

3. **Y a-t-il des contraintes en ce qui concerne la taille des données ?**

En général, les standards sont assez volumineux.

4. **Entrées / sorties ?**
5. **Quel type d'affichage des résultats souhaitez-vous ?**

Il faudrait pouvoir avoir différents formats en entrée.
En sortie, il faudrait pouvoir produire des schémas.

Il serait intéressant d'avoir un logiciel qui extrait les concepts importants et fait les liens entre eux.
C'est possible si la syntaxe est très simple.
On obtiendrait un réseau sémantique à raffiner, basé sur la fréquence des mots qui reflèterait les concepts importants et l'endroit où les trouver dans le texte.

6. **Le système d'exploitation a-t-il une importance ?**

Non.

7. **Quelle est la langue principale ?**

L'anglais principalement vu que les documents sont en anglais.
Ce serait aussi plus facile pour les collaborations et les conférences

6.3. Interview de Nicolas Mayer – 23 janvier 2007

1. **Quel est le type de données dont vous vous servez ?**
2. **Quelles sont vos pratiques en matière d'analyse de texte ?**

Au niveau de la méthode, on prend un référentiel sur la gestion des risques et on en fait une lecture intégrale. On recherche les concepts essentiels sur la gestion des risques en sécurité, c'est-à-dire les principes qui reviennent tout le temps et qui sont surtout caractéristiques du domaine. On obtient alors une liste de vocabulaires avec leurs définitions.
Ensuite, on applique à un autre référentiel. On cherche les mêmes concepts (sémantique), mais qui peuvent avoir une syntaxe différente. C'est plus complexe. On reprend tous les éléments textuels qui permettent de définir un concept donné. Cela permet de voir les points communs entre les concepts.
On pratique une approche top-down. **Mais il faut aussi avoir une approche bottom-up en analysant les exemples**. Cela permet de valider les définitions.

Au niveau logiciel, on utilise Word pour pouvoir souligner les termes et on met le tout dans un tableau excel.

On fait parfois l'union de certains concepts parce qu'ils représentent la même chose.
Et un même concept peut avoir deux niveaux de granularité.

On n'a pas toujours des définitions des termes. Il faut parfois chercher des explications dans le texte.

Un élément important est que tout le travail est fait de manière itérative et incrémentale. Par exemple, en testant un autre document, on peut s'apercevoir qu'un terme n'était pas intéressant.

L'objectif est de faire un langage de modélisation. On recherche donc les concepts statiques. Mais il y a aussi des phases dynamiques. On a aussi des éléments de processus qui ne sont pas statiques. (Ceci est caractéristique de mon projet de thèse)

On travaille aussi avec d'autres personnes qui vont relire pour valider.

La fréquence peut être une indication pour déterminer quel concept est propre au domaine, mais pas toujours. Il faut que ce soit plus fréquent que dans d'autres documents. => peut être un indicateur de si ce concept est caractéristique du domaine

Tris des données en fonctions de critères ? Pourquoi pas.

3. **Y a-t-il des contraintes en ce qui concerne la taille des données ?**

Il s'agit en général de documents assez gros.

4. **Entrées / sorties ?**
5. **Quel type d'affichage des résultats souhaitez-vous ?**

Au niveau des entrées, on a des documents PDF et Word.

Au niveau de l'affichage, il faudrait chaque mot avec sa fréquence et des liens vers sa localisation.
Il faudrait également des liens entre les concepts essentiels qui sont souvent associés.
Mais la terminologie étant très diversifiée, il faudra toujours une analyse manuelle.

Il faut aussi pouvoir personnaliser les analyses.

6. **Le système d'exploitation a-t-il une importance ?**

Windows.

7. **Quelle est la langue principale ?**

Pour le logiciel, en français et en anglais.
Pour les données, en anglais.

6.4. Interview de Anne Devos – 23 janvier 2007

1. **Quel est le type de données dont vous vous servez ?**
2. **Quelles sont vos pratiques en matière d'analyse de texte ?**

Le travail porte sur l'analyse des organisations en sciences sociales.
Il y a trois types de document textuel :
1) des interviews semi-directives ;
2) des documents d'entreprise : journaux, procès-verbaux de réunions ;
3) des observations participantes : on intègre une équipe et on travaille avec elle en ayant un intérêt plus particulier pour leurs tâches par rapport à un objet précis, ex. : un outil de gestion des connaissances : on analyse les réunions.
Ces observations sont mises sur fiche (1° traitement) et sont ensuite traitées comme les interviews.

La manière de traiter :
Tous les textes sont récoltés par rapport à un objet précis en lien avec une recherche. On travaille dans un cadre conceptuel théorique qui oriente le regard et les questions posées dans les entretiens.
Avant de relire les textes, on établit une grille de codage où sont définies les principales thématiques en fonction de ce qui est recherché. Ces thématiques se voient attribuer un code pour simplifier la consultation.
Il s'agit d'une grille ouverte. On lit le premier document et on lui applique la grille. Celle-ci va évoluer au fil des lectures.

Un logiciel pourrait nous aider à appliquer la grille au texte en transformant les codes en mots-clefs et en indexant le texte.

On travaille sur des blocs de texte, plutôt que sur des phrases ou des mots, vu la taille des corpus. Parfois, le sens est implicite. Une automatisation complète n'est donc pas possible.
Mais ce serait bien d'avoir un logiciel qui affiche le texte et la grille en parallèle.

Tout au long du travail d'analyse, la grille est étendue puis restreinte.
On procède de manière itérative en faisant 3-4 passages.
La grille de codage basée sur le cadre conceptuel et les questions devient descriptive, puis on opère des regroupements quand les codes sont proches.

Au point de vue outil, on ne sait pas toujours ce qu'on cherche. On fait des ajouts au fur et à mesure en fonction d'une heuristique propre.

Il y a un va-et-vient entre les questions, le cadre conceptuel et la grille de codage.

Un logiciel devrait pouvoir être nourri et permettre d'organiser clairement les codes.

Si on a des gros codes qui se raffinent, il est intéressant de pouvoir garder des liens entre les subdivisions.

L'objectif est d'avoir une vision transversale et d'aller vers une synthèse du corpus.
Automatiser complètement serait difficile. Par rapport à un bloc, on cherche l'élément, et ça, un logiciel ne peut pas le faire. Mais il peut aider à la construction de la grille de codage.

On travaille généralement sur des temps très longs. Il y a donc parfois des confusions dans les codes. Un logiciel qui aiderait à préserver ces codes serait utile.

On n'utilise pas d'analyses statistiques. La fréquence peut être utile, mais pas déterminante.
On fait de l'analyse qualitative. On ne code pas la fréquence de mots.
De plus, la terminologie n'a pas toujours de sens par rapport à l'analyse.

Dans le cadre de la démarche itérative, à un certain moment, on fait une synthèse des documents. On garde des citations exemplatives pour illustrer la recherche.
On sort le sens général du discours et on reprend les verbatims exemplatifs.
Après cette diminution des textes, on réapplique la grille pour avoir des éléments de comparaison. Cela permet de faire des regroupements dans la grille.
Ensuite, on réapplique cette dernière grille au corpus initial.

Au dernier codage, on garde les citations pertinentes, qui peuvent être différentes de celles de la synthèse.

Quand le travail doit être réalisé plus vite, on ne fait qu'un seul passage.

Au niveau du codage, s'il y a deux codes, c'est une indication que le code est trop général et on affine.
Puis, quand on a beaucoup trop de codes, on les synthétise.

Une fois que le corpus est codé, on procède par comparaison. On cherche les rapports entre le fonctionnement de l'organisation et la question de recherche.
On fait ces comparaisons au sein de groupes d'acteurs et entre groupes, et même entre études.

En ce qui concerne la fréquence, on peut voir si ça revient dans un groupe d'acteurs et si c'est corroboré dans d'autres groupes. On compare ce qui apparaît chez les uns et les autres, ou ce qui n'apparaît pas.

L'analyse de blocs se fait au niveau sémantique et par jeu de comparaisons. On cherche l'objectif énoncé par l'acteur, explicite ou sous-jacent. L'analyse se fait toujours dans le contexte.

Actuellement, tout se fait à la main. On retranscrit intégralement les interviews.
Tout est mis dans un tableau à deux colonnes avec le texte et le code en vis-à-vis.

Au niveau des synthèses individuelles, il y a des regroupements de code.
Au point de vue de l'analyse des comparaisons, on recherche des régularités entre des idées et des oppositions, plus que des fréquences.

Dans la grille, on met également la définition des codes.

Les grilles sont réutilisables. On peu les construire sur base de ce qui existe avant. Mais chaque chercheur a son « truc ».
On peut faire des échanges des entretiens et de la grille d'entretien.
Parfois, dans les travaux de grande ampleur, plusieurs personnes sont amenées à coder le même corpus.

Outil de classification de texte ? Pourquoi pas, mais ce n'est pas parce qu'une personne parle d'un sujet que c'est primordial pour elle.
Dans le cadre de la comparaison, on procède par assemblage.

Il est intéressant de pouvoir ouvrir de nouvelles pistes, de nouvelles perspectives.

3. **Y a-t-il des contraintes en ce qui concerne la taille des données ?**

On a plusieurs centaines d'interviews qui peuvent prendre 35 pages environ.

4. **Entrées / sorties ?**
5. **Quel type d'affichage des résultats souhaitez-vous ?**

On a principalement des documents word. Ce serait bien de pouvoir analyser des sites web (ceux des entreprises), mais les formats sont différents.

Au niveau affichage, c'est important de pouvoir visualiser : schémas, outil de visualisation des grilles de codage, couleurs, etc.

6. **Le système d'exploitation a-t-il une importance ?**

Windows. Le plus proche de ce qu'on utilise

7. **Quelle est la langue principale ?**

Le français principalement.