CHAPITRE VII

Caractérisation d'un texte dans un corpus : du quantitatif vers le qualitatif

Aperçu

La définition de d'indicateurs chiffrés et de mesures, pour passer de la description à la caractérisation, amène à réexaminer la question du corpus, cette fois en tant qu'il est les données de travail, et le référentiel effectif, pour les calculs. Tout ensemble de texte n'est pas un corpus, et différents critères ont été avancés : conditions de signifiance (*pertinence*, *cohérence*), conditions d'acceptabilité (*représentativité*, *régularité*, *complétude*), conditions d'exploitabilité (*homogénéité*, *volume*). A y regarder de plus près, le corpus s'articule en fait en plusieurs (sous-)corpus qui ont chacun un rôle dans l'analyse –*corpus existant*, *corpus de référence*, *corpus d'étude*, *corpus distingué*– dont on observe les modes de définition respectifs.

Les calculs sont un outil puissant et suggestif, ce qui ne dispense surtout pas d'en percevoir le fonctionnement et les limites. Pour comprendre et interpréter les formules proposées, pour en concevoir de nouvelles qui expriment d'autres propriétés, on se donne un modèle général pour la description des formules : identification des quantités mesurées (dans une notation structurée et unifiée), types des zones syntagmatiques, objectif de la mesure (pondérations, similarité). Cette boîte à outils est immédiatement mise en service pour interpréter des formules classiques, qui servent à illustrer des points de choix généraux : questions liées aux fréquences, aux variations de taille, aux emplois typiques de certains opérateurs (logarithme, moyennes,…).

Connaissant mieux les moyens des calculs sur les textes, la définition des unités caractérisantes s'appuie sur des propriétés textuelles mesurables et potentiellement significatives, dont on établit une liste aussi complète que possible. Le concept de *catégorie* permet d'articuler les informations quantitatives et qualitatives. Quant à l'évaluation des similarités entre textes, elle s'appuie sur le typage des unités (Communautés, etc.), qui gèrent les interrelations spécifiques entre unités de plus bas niveau. Les calculs s'opèrent sur des *profils d'approche*, qui favorisent l'efficacité du calcul et permettent un affinement progressif.

Table des matières du Chapitre VII

A. DÉFINIR UN CORPUS

1. Une question qui resurgit dans le contexte du calcul

Le corpus est nécessité et orienté par le traitement : c'est bien le préliminaire aux calculs, et c'est sous cet angle qu'il est considéré dans ce chapitre.

a) Les données

Le corpus se définit de fait comme l'objet concret auquel s'applique le traitement, qu'il s'agisse d'une étude qualitative ou quantitative.

> *corpus* : (ling.) ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique ; (lexicométrie) ensemble de textes réunis à des fins de comparaison, servant de base à une étude quantitative. (Lebart, Salem 1988, § *Glossaire*)

Mais les données ont un nom trompeur : elles ne s'imposent pas, elles sont construites. Certes, il y a un existant, directement sous forme de textes électroniques par exemple, –et donc l'analyste n'a pas une totale liberté d'« inventer » ses données, il part d'une réalité–, mais il reste des décisions du type : faut-il considérer tout ce qui est disponible ou en extraire un sous-ensemble plus significatif et équilibré ; comment tirer parti du codage disponible, comment éventuellement l'adapter au traitement envisagé. Le rapport aux données tient d'un compromis : faire avec ce à quoi on a accès, mais faire au mieux avec cela.

> La définition des textes et [le cas échéant des] fragments [qui subdivisent chaque texte] devrait dépendre du but de l'étude ; mais souvent, le statisticien ne peut qu'accepter les données disponibles... (Benzécri & *al.* 1981, p. 137)

Les linguistiques de corpus

L'accès actuel à de vastes ensembles de textes sous forme électronique a été une condition décisive pour le développement d'un courant linguistique récent : la linguistique à base de corpus (Habert, Nazarenko, Salem 1997).

L'approche à base de corpus revendique d'abord son réalisme, car elle se fonde sur des textes réels, des données attestées : le corpus s'oppose ici aux exemples *ad hoc* forgés pour les besoins d'une théorie ou d'une étude.

Le corpus est généralement l'apanage d'une linguistique descriptive, qui l'observe pour reconstituer *a posteriori* des régularités. Une linguistique normative peine à l'exploiter, car le corpus « brut » n'obéit pas au jeu de règles érigées *a priori*, si élaboré soit-il. Du côté des outils informatiques, le corpus appelle des traitements robustes, des analyses partielles.

b) Référentiel effectif

Le corpus fournit à la fois des éléments à étudier, mais aussi l'environnement descriptif de ces éléments. Le corpus est un tout, un vaste ensemble, qui constitue à lui seul le cadre et le référentiel de l'analyse Il met en présence les éléments, il fait qu'ils sont aussi considérés dans leur interrelation globale. Les éléments prennent alors une valeur relative par rapport au corpus : affinités et associations, fréquence ou rareté, banalité ou spécificité, etc.

Le cadre fixé par le corpus, souvent celui d'une application et d'une pratique, devient un moyen de réduire et d'ajuster l'appareil descriptif, grâce à un opportunisme efficace. On reprend et on adapte les ressources traditionnelles : ontologie et dictionnaire (limités au domaine), scripts (juste ceux associés aux situations envisageables dans la pratique concernée), lois de structuration du texte (sur la base de la forme conventionnelle du genre). Certains sombres problèmes des Traitements Automatiques des Langues trouvent soudain une issue : l'ambiguïté s'estompe, car dans un domaine fixé la langue prend un tour univoque ; l'implicite est dévoilé, puisque le corpus est ancré dans un cadre stéréotypé donné ; la granularité (ou niveau de détail) de la description trouve une juste mesure,

en fonction de la définition du corpus et de l'application envisagée. (Pincemin, Assadi, Lemesle 1996, §7.1) (Péry-Woodley 1995, §3)

2. Le corpus : un ensemble de textes ?

a) Tout ensemble de textes n'est pas un corpus : propriétés recherchées

Le corpus ne se laisse pas uniquement définir formellement, comme un ensemble de texte ou une suite de caractères alphanumériques. Il vérifie trois types de conditions : des conditions de signifiance, des conditions d'acceptabilité, et des conditions d'exploitabilité.

- *Conditions de signifiance* : Un corpus est constitué en vue d'une étude déterminée (*pertinence*), portant sur un objet particulier, une réalité telle qu'elle est perçue sous un certain angle de vue (et non sur plusieurs thèmes ou facettes indépendants, simultanément) (*cohérence*).
- *Conditions d'acceptabilité* : Le corpus doit apporter une représentation fidèle (*représentativité*), sans être parasité par des contraintes externes (*régularité*). Il doit avoir une ampleur et un niveau de détail adaptés au degré de finesse et à la richesse attendue en résultat de l'analyse (*complétude*).
- *Conditions d'exploitabilité* : Les textes qui forment le corpus doivent être commensurables (*homogénéité*). Le corpus doit apporter suffisamment d'éléments pour pouvoir repérer des comportements significatifs (au sens statistique du terme) (*volume*).

Chacune de ces conditions demande à être commentée, à partir des éclairages complémentaires, et assez remarquablement convergents, issus des différentes disciplines qui utilisent les corpus (statistiques lexicales et lexicométrie, analyse de contenu en psycho-sociologie, linguistique structurale, etc.).

Pertinence

Le corpus prend sens par rapport à un objectif d'analyse. Cela n'est pas sans incidence sur la question de sa réutilisabilité : à quelles conditions ce qui a été rassemblé pour servir un objectif peut être recyclé pour en servir un autre ? Une partie de la réponse se trouve dans l'explicitation des choix et conditions de recueil du corpus. D'autre part, ce n'est pas nécessairement le corpus tel quel qui est repris : le corpus original sert de source pour construire un autre corpus, dans le respect du nouveau contexte d'analyse.

> *Règle de pertinence* : Les documents retenus doivent être adéquats comme source d'information pour correspondre à l'objectif qui suscite l'analyse. (Bardin 1977, §III.I.1, p. 128)

Cohérence

L'analyse du corpus mène à une représentation synthétique, qui doit donc, pour être claire et expressive, pouvoir être comprise comme la représentation d'une entité, avec ses articulations internes et non comme la juxtaposition de plusieurs réalités indépendantes. C'est par le même geste, que l'on se donne un corpus, et que l'on s'isole de toutes les problématiques générales ou étrangères.

> Le caractère idiolectal des textes individuels ne nous permet pas d'oublier l'aspect éminemment social de la communication humaine. Il faut donc élargir le problème en posant comme principe qu'un certain nombre de textes individuels, à condition qu'ils soient choisis d'après des critères non linguistiques garantissant leur homogénéité, peuvent être constitués en corpus et que ce corpus pourra être considéré comme suffisamment isotope.
>
> [...] ce qui permet [par exemple] de réunir une cinquantaine de réponses individuelles en corpus collectif, c'est un ensemble de caractères communs aux testés : leur appartenance à la même communauté linguistique, à la même classe d'âge ; c'est aussi le même niveau culturel, la même « situation de testés ».
>
> (Greimas 1966, §VI.3, pp. 93-94)

> *Règle d'homogénéité* : les documents retenus doivent être homogènes, c'est-à-dire obéir à des critères de choix précis et ne pas présenter trop de singularité en dehors de ces critères de choix.

> Par exemple, des entretiens d'enquête, effectués sur un thème donné, doivent : être tous concernés par ce thème, avoir été obtenus par des techniques identiques, être le fait d'individus comparables. Cette règle est surtout utilisée lorsqu'on désire obtenir des résultats globaux ou comparer les résultats individuels entre eux.
>
> (Bardin 1977, §III.I.1, p. 128)

> Lorsque nous utilisons [le terme *corpus*], nous sous-entendons 'corpus de documents homogènes', à savoir un ensemble de documents qui ne soit pas hétéroclite. Il ne s'agit pas de considérer n'importe quel ensemble de documents sans aucun rapport les uns avec les autres. Par exemple, un ensemble de brevets relatifs aux céramiques, un ensemble de publications mondiales sur l'intelligence artificielle constituent pour nous, des corpus homogènes. Les traitements que nous exposerons par la suite sont envisagés sur de tels corpus. (Chartron 1988, §II.1, p. 16)

> Le choix d'un corpus présuppose... que ce corpus constitue bien un *objet d'étude* ; c'est-à-dire, que l'analyste le perçoive comme une entité ou un *objet* dans l'univers référentiel qui l'intéresse. En définitive, même si ce n'est que de manière implicite, l'analyste fait des hypothèses sur les conditions d'existence de cet objet, sur ses lois de production, sur les paramètres qui le font reconnaître dans cet univers référentiel. (Reinert 1990, §1.2, p. 27)

Représentativité

Les statisticiens soulignent bien que définir un échantillon est une opération complexe, pour assurer que l'extrait présente la même configuration des observables. La réalité à décrire présente un certain équilibre, une certaine composition, que le corpus doit d'efforcer de refléter.

> *Règle de représentativité* : On peut, lorsque le matériel s'y prête, effectuer l'analyse sur échantillon. L'échantillonnage est dit rigoureux si l'échantillon est une partie représentative de l'univers de départ. Dans ce cas les résultats obtenus sur échantillon seront généralisables à tout l'ensemble.
>
> Pour échantillonner il faut pouvoir repérer la distribution des caractères des éléments de l'échantillon. Un univers hétérogène demande un échantillon plus important qu'un univers homogène. [...] Comme pour un sondage, l'échantillonnage peut se faire au hasard, ou par *quotas* (les fréquences des caractéristiques de la population étant connues, on les reprend dans des populations réduites pour l'échantillon).
>
> (Bardin 1977, §III.I.1, p. 127)

Pour la linguistique, ce qui autorise des études sur des corpus toujours limités, c'est la nature redondante de la langue et la clôture des unités textuelles.

> Le corpus n'est [...] jamais que partiel, et ce serait renoncer à la description que de chercher à assimiler, sans plus, l'idée de sa représentativité à celle de la totalité de la manifestation. Ce qui permet de soutenir que le corpus, tout en restant partiel, peut être représentatif, ce sont les traits fondamentaux du fonctionnement du discours retenus sous les noms de *redondance* et de *clôture*. Nous avons vu que toute manifestation est itérative, que le discours tend très vite à se fermer sur lui-même : autrement dit, la manière d'être du discours porte en elle-même les conditions de sa représentativité. (Greimas 1966, §IX.1.b, p. 143)

Quand l'étude vise à décrire la langue ou le fonctionnement des textes « en général », la condition de représentativité semble devoir se traduire par une recherche de diversité maximale. Autrement dit, dans l'idéal, tous les cas de figure existants doivent être présents dans le corpus. Deux tactiques sont observables : la course à la quantité d'une part (engranger le maximum de données, le poids total devant être garant de la richesse amassée), la construction raisonnée d'autre part (se donner une grille quadrillant la réalité, et s'en servir pour rassembler méthodiquement des textes correspondant à tous les aspects recensés). La première tactique, dont la devise est « more data is better data » (Péry-Woodley 1995, §2.3.1), est manifestement grossière, mais souvent elle est justifiée (en partie) par les difficultés profondes auxquelles se heurte de plein fouet la seconde tactique : quel modèle adopter pour organiser la sélection des textes, qui ne porte pas sa part d'*a priori* réducteurs ? Plus gravement, la problématique elle-même apparaît utopique irréaliste : il n'y a pas de langue

générale, ou standard, ou moyenne ; et les textes sont tous pris dans des pratiques qui les contextualisent[1].

> La recherche de corpus équilibrés semble bien constituer une impasse : la notion d'équilibre s'apparente à celle de « langue générale », et elle paraît tout aussi insaisissable. Elle suppose également une recherche irréaliste d'exhaustivité : le corpus équilibré est sans doute celui qui a « de tout un peu », mais encore faudrait-il savoir ce qu'est « tout », c'est-à-dire quelles sont les classes à représenter, –ce qui nécessite un modèle complet de la variation –, et avoir accès à des textes les représentant. (Péry-Woodley 1995, §2.3.2, p. 218)

Admettre la relativité et la part de choix qu'il y a dans la constitution de tout corpus, c'est également reconnaître le caractère décisif de l'établissement du corpus. En particulier, bien souvent le corpus (ou une de ses parties) est utilisé comme référentiel (puisqu'il est représentatif de la réalité à décrire) et il conditionne tous les résultats de l'analyse.

> Le choix d'une norme endogène au corpus, le tout comme étalon des parties, est justifié par le fait maintenant bien établi qu'une forme [i.e. une unité], quelle qu'elle soit, n'a pas de fréquence en langue. (Note : Certains auteurs, contre toute évidence, affirment le contraire et invoquent des probabilités de langue. En revanche, nous sommes bien conscients du fait que l'usage d'une norme intrinsèque confère à l'élaboration du corpus une écrasante responsabilité.) (Lafon 1980, p. 137)

Régularité

La régularité correspond au fait que l'on explicite des principes pour définir le corpus, sans se permettre d'exceptions qui introduiraient des écarts locaux (manques, excès, éléments étrangers).

> *Règle de l'exhaustivité* : une fois défini le champ du corpus (entretiens d'une enquête, réponses à un questionnaire, éditoriaux d'un quotidien de Paris entre telle et telle date, émissions de télévision concernant tel sujet, etc.), il faut prendre en compte tous les éléments de celui-ci. Autrement dit, il n'y a pas lieu de laisser un élément pour une raison quelconque (difficulté d'accès, impression de non-intérêt) non justifiable sur le plan de la rigueur. Cette règle est complétée par la règle de non-sélectivité.
>
> Par exemple, on réunit un matériel d'analyse des publicités pour automobiles parues dans la presse pendant une année. Toute annonce publicitaire répondant à ces critères doit être recensée.
>
> (Bardin 1977, §III.I.1, p. 127)

> [Exigence d']exhaustivité : les ensembles [des individus et des variables] représentent un inventaire complet d'un domaine réel dont le cadre n'est guère discutable. (Benzécri & *al.* 1973b, §A.2.1.3, p. 21)

Complétude

Le corpus doit avoir un niveau de détail adapté aux besoins de l'analyse : les adaptations nécessaires peuvent être soit de l'enrichir et de l'affiner, soit d'ajuster, par réduction, le niveau de discrétisation de la réalité à représenter réalisée à partir des données.

> L'exhaustivité du corpus est [...] à concevoir comme l'adéquation du modèle à construire à la totalité de ses éléments implicitement contenus dans le corpus. (Greimas 1966, §IX.1.b, p. 143)

[1] Une voie envisagée a donc été de s'appuyer sur une description systématique des situations de communication et de production des discours. On se donne un ensemble de paramètres, tels que : la communication directe (interlocution) ou différée, l'adresse à un public/lectorat collectif ou non, le caractère formel de l'échange, etc. C'est la méthode adoptée dans (Bronckart & *al.* 1985). Douglas BIBER (Biber 1988) recule d'un cran le caractère nécessairement subjectif d'une telle grille, en se fondant non pas directement sur les pratiques de communication (et donc les genres), mais en partant d'un ensemble de caractéristiques linguistiques (essentiellement morpho-syntaxiques) pressenties comme liées à la diversité des genres. L'étude dépend donc toujours, mais cette fois-ci indirectement, d'une certaine perception que l'on a des genres. Même si la statistique (analyse factorielle) a un pouvoir certain de généralisation (gommage d'éléments non pertinents, interpolation à partir d'un nombre limité d'éléments, caractère suggestif des représentations), les résultats de Douglas BIBER doivent être compris comme relatifs aux choix initiaux (textes utilisés pour l'étude, choix des traits morphosyntaxiques représentatifs).

> *exhaustivité* : l'exhaustivité des données (qui assure à l'analyse une base intrinsèque [...]) peut, conformément au principe d'équivalence distributionnelle, être assurée par une partition [...], ou [par le] choix d'un échantillon fini (éventuellement stratifié [...]) sur un espace potentiel continu (Benzécri & *al*. 1973, § *Indice systématique*)

Homogénéité

Sachant l'objectif de l'analyse, et les dimensions de variation que l'on veut étudier, le corpus doit être aussi homogène que possible pour ses autres caractéristiques.

> [Exigence d']homogénéité : toutes les grandeurs recensées [...] sont des quantités de même nature. (Benzécri & *al*. 1973b, §A.2.1.3, p. 21)
>
> *homogénéité* : pour définir objectivement le tableau des données étudiées [...], on vise à l'homogénéité des variables : ce qui permet l'adoption d'une unité de mesure unique [...] ; l'homogénéité est autorisée par l'hypothèse du *nexus*, [à savoir celle de l'] interrelation de tous les caractères d'un vivant (Benzécri & *al*. 1973, § *Indice systématique*)

Volume

Les procédés d'analyse visent à saisir et décrire des régularités qui structurent le corpus. Une certaine redondance est nécessaire pour que puissent émerger et être repérés des aspects caractéristiques et informatifs.

> Le logiciel ALCESTE est un outil d'aide à l'interprétation d'un corpus textuel : entretiens, réponses à une question ouverte, textes littéraires, en fait tout document écrit à l'aide de l'alphabet latin, des dix chiffres et des signes usuels de ponctuation pourvu qu'il présente une certaine homogénéité et un volume minimum. [...]
>
> Il y a toutefois deux conditions pour obtenir un résultat signifiant : la première est que le corpus présente une certaine cohérence thématique [cf. condition d'*homogénéité*]. C'est le cas (en général !) des réponses à une question ouverte, de textes littéraires, de recueils d'articles sur un sujet, etc... *A contrario* on ne peut pas espérer une indication de contenu pour un patchwork de fragments disparates, aussi intéressants soient-ils isolément...
>
> La seconde est que le document soit suffisamment volumineux pour que l'élément statistique entre en ligne de compte. C'est du reste l'intérêt d'ALCESTE de donner très rapidement une vision globale sur une documentation volumineuse qui serait autrement très longue à dépouiller.
>
> (Reinert, Piat 1995, cahier 1, §0, p.3)

La condition de volume est importante pour des analyses statistiques, pour que celles-ci puissent être considérées significatives. En revanche, présenter la recherche de volume essentiellement comme un moyen d'obtenir une bonne représentativité 'générale' (Church & Mercer 1993) est déplacé : le volume et la représentativité sont des caractéristiques à part entière, complémentaires.

Dans le cas d'une exploitation manuelle, c'est-à-dire sans l'outil informatique, on s'inquiétera à l'inverse de la *maniabilité* du corpus (Garcia-Debanc 1989, p. 44).

b) Du texte, des textes

Certains travaux ne considèrent pas les unités que forment les textes, ils ne visent que le matériau linguistique, à savoir une seule des facettes du texte. Le corpus est alors un ensemble de données pour des études de la langue.

> Nous employons le mot *corpus* dans une acception restreinte empruntée à J. Sinclair [...] : « Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage. » (Habert, Nazarenko, Salem 1997, p. 11)

C'est particulièrement à Marie-Paule Péry Woodley que l'on doit d'avoir questionné le bien-fondé de corpus linguistiques mais non textuels, qui rassemblent *du* texte et non *des* textes. En effet, ce choix s'apparente à un appauvrissement systématique et injustifié : toute manifestation linguistique ressort d'une forme de textualité, au sens d'une unité de communication. Et toute suite linguistique reçoit une part de sa définition et de sa significativité de son contexte textuel (par le genre auquel se rattache le texte, ou encore par le jeu des zones de localité, du syntagme à l'étendue du texte entier). S'en tenir à des extraits, même larges, ne suffit pas à rendre compte du fonctionnement global

qu'institue l'unité texte. De plus, le corpus devient alors un bloc monolithique et inerte, car les dimensions qui auraient permis sa redéfinition pour une autre étude ont été effacées.

> Notons l'absence d'article devant le mot *text* dans la phrase de K. Church et R. Mercer citée plus haut [...] : il s'agit d'analyser *du* texte et non *des* textes. Se pose ici la question de la pertinence de l'unité texte dans la constitution et l'analyse de corpus : un ou des textes par opposition à du texte. Un corpus se compose par définition de discours, de langue concrète [...], et c'est immanquablement sous la forme de textes –écrits ou parlés– que la langue se réalise en discours. Cette langue concrétisée en textes porte les marques des conditions de leur production, et des objectifs qui les ont motivés. A l'extrême, recueillir du texte, c'est ne se donner aucun moyen d'échantillonnage, c'est soumettre ensuite à l'analyse un objet dont l'hétérogénéité est totalement opaque, c'est enfin se priver de toute possibilité de prise en compte de la structure textuelle. Les conséquences d'un tel choix se situeraient donc dans l'immédiat sur le plan de la qualité du travail possible sur un tel corpus, et au delà sur le plan de sa réutilisabilité. (Péry-Woodley 1995, §2.3.3, pp. 218-219)

3. Constitution : une typologie des corpus en présence

a) Emboîtements

En prenant le mot corpus dans son sens le plus large, il s'avère que l'analyste n'a pas affaire à un corpus –un ensemble de textes–, mais à une série de corpus[2], qui ont chacun leur rôle.

- *Le corpus existant (ou corpus latent)* : l'ensemble des textes auxquels on peut avoir accès, dont on peut disposer. C'est généralement une masse « informe », non systématique, mal défini, aux contours incertains. Il est difficile d'en avoir une vue globale. Cet existant dépend de conditions étrangères à l'étude, qui ne sont pas toutes connues ni maîtrisées.
- *Le corpus de référence* : il est composé à partir du corpus existant, en adéquation avec l'objectif de travail ; il est clairement défini et équilibré. C'est lui qui fournit l'univers le plus large dans lequel chaque élément trouve sa valeur. Il constitue l'univers et fixe le point de vue de l'étude. Il représente le fond sur lequel on veut profiler les textes étudiés. Autrement dit, il matérialise le contexte, actif et virtuel, et acquiert là son statut linguistique d'unité de rang supérieur –la linguistique ne s'arrêtant ni à la phrase, ni même au texte (Rastier 1998, §III.2).
- *Le corpus d'étude* : c'est l'ensemble des textes sur lesquels porte effectivement l'analyse, pour lesquels on attend des enseignements, des résultats. Le corpus d'étude n'est pas nécessairement une partie du corpus de référence, mais le corpus de référence doit pouvoir être considéré comme représentatif du corpus d'étude, pour l'aspect dont on veut rendre compte. Paradoxalement, le corpus d'étude peut être plus volumineux que le corpus de référence : ce qui est définitoire, ce n'est donc ni un rapport d'inclusion, ni un rapport de taille, mais la spécificité des rôles de chacun.
- *Le corpus distingué* : c'est un groupe de textes du corpus d'étude que l'on veut caractériser dans leur cohésion d'ensemble, par rapport au reste du corpus d'étude.

Exemples illustratifs, d'après des travaux actuels :

[2] On pourrait aussi préférer réserver le terme *corpus* pour les ensembles de textes rassemblés pour l'analyse et qui en fournissent le contexte (*corpus de référence*, *corpus d'étude*). Le *fonds documentaire* désignerait alors l'ensemble des textes à disposition (de préférence à *corpus existant*) ; quant au *corpus distingué*, il correspond habituellement à ce que l'on appelle *sous-corpus*.
Le choix ici de décliner quatre *corpus*, malgré leurs différences profondes, se justifie par l'intention de souligner les usages contrastés du mot *corpus*.

	corpus existant	**corpus de référence**	**corpus d'étude**	**corpus distingué**
Etude d'Etienne Brunet (Brunet 1995)	la base Frantext de l'INaLF	350 romans entre 1830 et 1970	phrases de ces romans comportant au moins une des 165 unités lexicales retenues pour définir la thématique du sentiment	les éléments retenus dans les romans d'un romancier
Construction des profils pour l'application DECID de diffusion ciblée	textes enregistrés dans la base SPHERE de la DER d'EDF, autres textes électroniques collectés de façon centralisée.	l'ensemble des textes d'Action, en version définitive, à partir de l'année 1990 jusqu'à l'année en cours.	les textes d'Action pour une année (le cas échéant, les textes en version provisoire pour l'année suivante).	les textes d'Action du corpus d'étude, dont le rédacteur (plus exactement le responsable) est rattaché à un Département donné.

Chaque choix est significatif, et joue un rôle pour la suite de l'analyse. Par exemple, Greimas montre l'incidence de ce qui est pour nous le corpus de référence :

> [Pour l'étude de l'univers de Bernanos,] la question pratique [...] est de savoir quelle signification il faut attribuer respectivement aux trois corpus possibles : le corpus ayant les dimensions d'un roman, le corpus de la totalité des écrits de Bernanos et, enfin, le corpus de tous les romans d'une société et d'une période historique données, et quelles corrélations structurelles on peut raisonnablement espérer retrouver entre les modèles qu'on pourra expliciter à partir de tels corpus.
>
> [...] d'une part, les corpus constitués par des romans-occurrences sont à considérer comme des inventaires de modèles implicites permettant la construction du genre « roman du XX^{ème} siècle » ; [...] d'autre part, les corpus faits de totalités représentatives de paroles individuelles constituent autant de manifestations pouvant servir à la construction d'un genre désigné sommairement comme « style de la personnalité » [...].
>
> Un roman-occurrence, le *Journal d'un curé de campagne*, [...] se trouve placé au croisement de deux axes, et [est] susceptible d'entrer simultanément dans deux corpus différents et d'être soumis à deux analyses ayant des visées divergentes. Pour ne prendre, à titre d'exemple, que l'analyse actancielle, on voit que les personnages de ce roman pourront être considérés comme les variables d'une structure actancielle romanesque propre à la littérature du XX^{ème} siècle, mais qu'ils participeront en même temps, comme des incarnations spécifiques, de la structure actancielle proprement bernanosienne.
>
> (Greimas 1966, §IX.1.f, pp. 148-149)

Une telle explicitation des articulations entre les différents ensembles de textes à considérer, chacun avec leur rôle dans l'étude, a déjà fait l'objet de réflexions, en Analyse du Discours par exemple :

> Nous introduisons [...] trois concepts complémentaires, ceux d'*univers discursif*, de *champ discursif* et d'*espace discursif*.
>
> On entendra par « univers discursif » l'ensemble des énoncés de tous types qui coexistent, ou plutôt interagissent, dans une conjoncture. Cet ensemble est nécessairement fini, mais irreprésentable, jamais pensable dans sa totalité par l'AD [Analyse du Discours]. Quand on utilise cette notion, c'est essentiellement pour y découper des « champs discursifs ».
>
> Le « champ discursif » est définissable comme un ensemble d'archives qui se trouvent en relation de concurrence, au sens large, et se délimitent donc pour une position énonciative dans une région donnée. Le découpage de tels champs doit découler d'hypothèses explicites et non d'une partition spontanée de l'univers discursif. Certes, la tradition a légué un certain nombre d'étiquettes (champs discursifs religieux, politique, littéraire, etc.), mais ce sont là des grilles extrêmement grossières, de peu d'intérêt pour l'AD, qui est contrainte à prendre en compte de multiples paramètres pour construire des champs pertinents.
>
> L'« espace discursif », enfin, délimite un sous-ensemble du champ discursif, lie au moins deux archives dont il est permis de penser qu'elles entretiennent des relations privilégiées, cruciales pour la compréhension des discours concernés. C'est donc une décision de l'analyste qui le définit, en fonction de ses objectifs de recherche. Si on découpe de tels sous-ensembles, ce n'est pas par simple

> commodité (parce qu'il serait difficile d'appréhender un champ discursif dans sa totalité) mais aussi et surtout *parce qu'une archive donnée ne s'oppose pas de manière semblable à toutes celles qui partagent son champ* : certaines oppositions sont fondamentales, d'autres ne jouent pas directement un rôle essentiel dans la constitution et la préservation de l'archive considérée.
>
> Aucun champ discursif n'est insulaire ; il existe une circulation intense d'une région à une autre de l'univers discursif, mais les voies qu'elle emprunte n'ont rien de stable ; selon les discours et les conjonctures concernés on aura affaire à des jeux d'échanges très différents. [...]
>
> Cette étude des échanges entre champs débouche immédiatement sur la question de l'*efficacité* des discours, de leur aptitude à susciter l'adhésion d'un ensemble de sujets. Le réseau de renvois d'un champ à l'autre (qu'il s'agisse de citations explicites, de schèmes tacites, de captations,...) ne contribue pas peu à cette efficacité : confronté à un discours de tel champ, un sujet retrouve des éléments élaborés ailleurs qui, en intervenant subrepticement, créent un effet d'évidence. On assiste à une « métaphore », un transport généralisé d'un champ à l'autre (mais pas de n'importe quel champ à n'importe quel autre) sans qu'il soit possible de définir un lieu d'origine, un sens « propre » ; tout simplement parce que la question même de l'origine n'est pas pertinente ici.
>
> (Maingueneau 1991, §4.3, pp. 158-159)

Il ne paraît pas abusif d'établir la correspondance suivante, même si la superposition des deux modèles n'est pas totale (par exemple, chez Maingueneau, l'espace discursif est inclus dans le champ discursif) :

Maingueneau :	univers discursif,	champ discursif,	espace discursif.
Pincemin :	corpus existant,	corpus de référence,	corpus d'étude.

Les travaux dans le domaine font également souvent état de sous-corpus. Ce qui est appelé *sous-corpus* est tantôt un corpus d'étude (pris comme une partie du corpus de référence), tantôt un corpus distingué (détaché du corpus d'étude). Le sous-corpus mérite une attention particulière, en ce qu'il est davantage qu'un corpus : non seulement, comme son nom l'indique, il entretient une relation privilégiée avec le corpus dont il est extrait ; mais aussi, sa nature est différente –il n'est pas toujours un ensemble de textes.

C'est cette question qui est examinée dans les paragraphes qui suivent, après une discussion sur le mode de contextualisation opérée par le corpus de référence.

b) L'intertexte : le corpus comme contexte et comme totalité

Déterminer le contexte est bien sûr un acte herméneutique majeur, puisque c'est décider ce qui est accessible, et même structurant, pour l'interprétation du texte, *en dernière instance*.

> L'extension de corpus.
>
> [...] le contexte sémantique d'un sémème n'a d'autres limites que celle du texte ; [...] les relations sémiques d'afférence peuvent excéder le contexte syntaxique, et relier des sémèmes en n'importe quel point du texte, avec un effet cumulatif ; cela est particulièrement clair avec les noms propres. Ce type d'extension repose toutefois sur une hypothèse forte : que le texte empirique est partout identique à lui-même dans la mesure où il mettrait partout en œuvre les mêmes types de systématicité.
>
> On peut aussi étendre le corpus à l'ensemble de l'œuvre du même auteur. Selon une remarque (incidente) de Hjelmslev (1973, p.151), l'œuvre d'un auteur est la plus grande unité linguistique possible. Même dans le cas –privilégié– où l'on connaît l'auteur d'une série de textes, cette affirmation repose elle aussi sur une hypothèse forte : l'identité à soi de l'auteur –entendu comme idiolecte.
>
> (Rastier 1987, §IX.4.2.1, p. 252)

Il s'agit d'un tout par rapport auquel se définit chaque texte, chaque élément du corpus étudié ; or, la linguistique nous avertit de trois totalités illusoires :

> il faut abandonner trois totalités romantiques, séduisantes, mais infondées, sinon dans une ontologie : *(i)* Celle du texte [isolé] [...]. La notion de « clôture textuelle » chez les contemporains doit beaucoup à cet unitarisme romantique [...]. *(ii)* Celle de l'œuvre, à laquelle répond la notion de style individuel [...]. *(iii)* Celle de l'Intertexte, qui dérive de la notion schlegelienne de totalité littéraire. [...] Il n'est même pas exclu qu'aujourd'hui l'Hypertexte soit le dernier avatar de la Totalité romantique des textes. (Rastier 1998, §III.2, pp. 107-108)

Autrement dit, le texte ou l'œuvre sont des unités, que l'on peut étudier en tant que telles, mais non pas des totalités « définitives ». Si par exemple on étudie les romans d'un auteur isolément, il faut avoir conscience que l'on fait abstraction d'une dimension significative, la « profondeur » qui

les met en relief dans un contexte intertextuel, comme l'appartenance à un genre. En revanche, ces unités sont d'excellents candidats au *corpus distingué*, et peuvent concourir, par un cheminement inverse d'extension (vs de focalisation), à la définition du corpus de référence.

> A partir d'un texte [note : Nous convenons que le texte permet de recruter son intertexte, cf. (Rastier 1989, § 2)], l'intertexte est ce par quoi l'on accède par l'ensemble des références (ou allusions) et plus généralement par l'ensemble des connexions opérées par la lecture et qu'on peut appeler l'*anagnose* [, selon la définition de Ioannis Kanellos et Théodore Thlivitis]. (Rastier 1998, §III.2, p. 108)

> Un texte n'est [pas] interprété « hors-contexte » mais au sein d'un *univers de textes*, que nous appelons *anagnose* et qui porte la trace d'une intention interprétative. [...]
>
> Un texte peut [...] appartenir à plusieurs intertextes, instanciant ainsi à sa mesure, différents points de vue. [...] l'intertexte constitue une sélection du lecteur, effectuée selon ses propres objectifs interprétatifs, et qui sert à « soutenir » les relations sémantiques qu'il désire mettre en avant.
>
> (Thlivitis 1998, §1.3 & 2.1.3, pp. 29 & 41)

Le *genre* est, lui, un candidat au corpus de référence, ou du moins un paramètre important de sa construction. Il rallie en effet le texte à une situation dans la réalité, dans les pratiques de rédaction et de lecture.

> Dans la problématique du texte, le contexte, contrôlé par le texte, se décline en zones de localité. Les éléments pertinents de la situation sont requis par l'analyse du texte : tout texte, par son genre, se situe dans une pratique. Le genre est ce qui permet de relier le contexte et la situation, car il est à la fois un principe organisateur du texte et un mode sémiotique de la pratique en cours. [...]
>
> Le texte semble certes en linguistique une unité maximale. Mais un point de vue plus philologique engage à considérer que l'ensemble des textes relevant d'un même genre (et d'une même langue) constitue un « bon » corpus au sein duquel il est possible de caractériser et d'analyser un texte. [note : le corpus est la seule objectivation possible (philologique) de l'intertexte, qui sinon demeure une notion des plus vagues.]
>
> (Rastier 1998, §III.2, p. 107)

Il faut toutefois garder leur autonomie aux deux concepts de *corpus* et de *genre*.

Les genres sont déterminés par les pratiques sociales. Ils sont reconnus et décrits par la linguistique, car c'est une réalité intertextuelle, par laquelle peuvent s'expliquer certaines affinités et certaines régularités entre des textes. Mais, notamment pour les besoins du codage et de la structuration des grandes bases textuelles, le genre court toujours le risque d'une définition théorique figée. Celui qui veut étiqueter et classer les textes d'un corpus par genre risque de voir la délimitation de ses « paquets » se dérober. Où commence et où finit le genre ? Parions que les discussions soient encore pour longtemps ouvertes.

Le corpus, lui, relève d'un point de vue, contingent, –parmi une multiplicité d'autres points de vue possibles[3]–, donnant un cadre à la constitution d'un objet. C'est un paramètre méthodologique, car l'étude veut que l'on se donne un domaine d'action, relatif à une recherche particulière[4]. Le corpus s'inscrirait davantage du côté de l'herméneutique que de celui de la linguistique. Il est défini par un objectif de lecture, d'analyse. Toute étude (de statistique textuelle, de texte) tôt ou tard,

[3] Sur la multiplicité des corpus / intertextes, voir notamment (Thlivitis 1998) : §1.2.2, p. 22 sq. (intertexte centré texte, ou centré auteur et plutôt descriptif, ou centré lecteur et plutôt productif) ; §2.1.3, p. 39 (incidence de l'intertexte sur ce que l'on perçoit dans la lecture d'un texte), et p. 41 (l'intertexte comme point de vue).

[4] Le caractère subjectif et singulier d'un corpus peut être relativisé, dans la mesure où il conduit à une exploitation et à des résultats s'inscrivant dans un cadre d'analyse commun et réutilisables. Voir par exemple l'effort de (Thlivitis 1998) dans ce sens, pour la réutilisation d''interprétations' (classes sémantiques) basées sur un corpus :

« la méthodologie de la Sémantique Interprétative Intertextuelle [...] oblige à l'explicitation des *sources sémiques* à l'origine de la constitution d'une classe sémantique. Le lecteur est donc indirectement incité à réutiliser les interprétations existantes, en y apportant sa propre interprétation. De cette manière, nous proposons le dépassement, à la fois, de la volonté d'atteindre une juste *objectivité descriptive* et de la liberté d'une *subjectivité descriptive* totale en les remplaçant par un *consensus inter-subjectif*, issu de l'interaction *multi-utilisateur* avec un *espace commun d'analyses* et soutenu par une méthode de travail interprétatif qui incite à la *consultation* et à la *réutilisation* récurrentes. » (Thlivitis 1998, §1.2.2.2, p. 28)

explicitement ou implicitement, se donne un corpus, dans lequel elle va piocher, calculer, contraster. Le corpus a été fixé, la manière dont cela a été fait peut être discutée, argumentée, mais de toutes façons la définition du corpus est effective.

En somme, un ensemble de textes relevant d'un même genre peuvent constituer un corpus intéressant et fructueux. Pour autant, ce n'est pas le seul corpus valide. Le genre est un facteur qui contribue à l'homogénéité du corpus, mais d'autres modes de cohésion sont possible.

c) Le sous-corpus est encore des textes

Eléments factuels

Un (sous-)corpus peut être constitué comme la réunion des textes qui ont un caractère factuel commun. Dans les bases bibliographiques, les caractères factuels sont prévus par des champs. On considère ainsi les œuvres de :

- un auteur,
- un genre,
- une période,

ces critères pouvant être croisés (œuvres d'un genre à une époque, œuvres d'un auteur dans un genre, etc.)

On focalise ainsi explicitement l'étude, évitant des mélanges et des hétérogénéités qui brouilleraient l'analyse. En termes de normes linguistiques, différents degrés de régularités sont ainsi observables : la période reflète un *dialecte*, le genre un *sociolecte*, l'auteur un *idiolecte*.

> Dans l'analyse thématique, [le corpus] doit être restreint à bon escient pour pouvoir caractériser la spécificité des discours et des genres : les thèmes du roman ne sont pas ceux de l'essai ni du poème. Ainsi, en dépouillant un corpus trop étendu qui mêlait des romans et des essais dans la période 1830-1870, nous nous sommes aperçu que les sentiments du roman n'étaient pas ceux de l'essai. Par exemple, le sentiment de fraternité, récurrent dans les ouvrages de Leroux, et celui d'équité chez Proudhon, n'ont pas été relevés dans les romans, à l'exception confirmatrice des Misérables, qui alterne des chapitres romanesques et d'autres qui relèvent du genre de l'essai.
>
> A supposer même que le mot se rencontre dans des genres différents, rien n'assure qu'il se rapporte aux mêmes thèmes : *amour* se rencontrera certes en poésie et dans le roman, mais le thème de l'Amour diffère pourtant avec ces genres. Il n'a pas la même molécule sémique, ni les mêmes lexicalisations, ni les mêmes antonymes ».
>
> (Rastier 1995a, §II.1, p. 235)

Ce mode de construction est plus complexe qu'il n'y paraît. Rassembler l'œuvre d'un auteur : jusqu'où aller ? faut-il considérer telle œuvre secondaire, tel brouillon, tel écrit atypique,... ? Si le critère est le genre : comme il n'y a pas de consensus sur la liste et le contenu des genres, est-ce que par exemple ce qui est déclaré comme roman ou comme poésie correspond à ce que l'on veut étudier ? (le roman par lettres est-il un roman, le poème en prose fait-il partie de la poésie ?). La période suppose elle aussi un travail de définition : notre culture raisonne par siècles ; mais ce découpage n'est-il pas arbitraire et brutal ?

> L'analyse qui cherche à expliciter les transformations diachroniques des structures ne doit pas utiliser le découpage du texte [corpus] en tranches, correspondant aux « pleines » réalisations des structures, mais opérer, au contraire, une division du texte en séquences superposées, comprenant chaque fois, des deux côtés de la zone franche, deux zones d'enchevêtrement où les structures survivantes coexistent avec les structures de remplacement nouvellement élaborées (Greimas 1966, §IX.1.g, p. 151)

Morceaux choisis : les textes dans les textes

On peut choisir d'étudier les divisions structurelles d'un texte comme autant de « textes » autonomes : chapitres (pour le roman), actes (pour le théâtre), etc. Considérer que l'on obtient encore des textes, c'est envisager le caractère fractal de la textualité : en délimitant un passage, l'auteur lui compose un début, une fin, une cohérence propre ; et chaque texte semble receler en puissance une multitude d'autres textes plus petits.

Un passage peut être délibérément isolé, lu et étudié pour lui-même. La décision de le définir est un acte herméneutique, qui décide de lui reconnaître et de lui assigner les propriétés d'un texte. Les recueils de « textes » et les anthologies sont bien des collections de tels passages.

> Nous dirons qu'un énoncé, un texte ou un corpus est un ensemble de mots ; un ensemble ordonné, obéissant aux lois d'un idiome, et dont la suite naturelle est porteuse d'une signification. On convient d'autre part de nommer *texte* tout énoncé ou toute succession dénoncés, tout discours ou fragment de discours, sans limitation d'étendue, provenant d'un même locuteur ou scripteur et présentant une certaine unité ; une collection définie de textes sera nommée *corpus*. (Muller 1977, p. 5, §1)

d) Le sous-corpus est du texte (réunion d'extraits)

Contextes d'un concept pôle

Le concept consiste en un mot ou un ensemble de mots : une thématique pressentie comme importante, ou encore les désignations d'un personnage principal (héros)[5]. Les contextes sont alors établis, par une condition de proximité plus ou moins élaborée : fenêtre (en nombre de mots), zone typographique (phrase délimitée par la ponctuation forte, paragraphe), construction syntaxique (par exemple adjectifs qui qualifient le nom pôle). Bien entendu, toute la suite de l'étude est orientée par ces choix initiaux : choix des pôles surtout, et choix du mode de détermination des contextes[6]. Le statut d'un pôle peut également varier, d'objet central et référentiel de l'étude, à prétexte provisoire ou amorçage.

> Si l'on veut préciser encore le rapport de l'analyse lexicale à l'analyse thématique, il faut préciser que le mot à partir duquel commence la recherche n'en est pas l'objet, à la différence d'un mot-vedette qui ferait l'objet d'une recherche lexicographique. On va certes chercher, en utilisant les moyens d'assistance informatisés, d'autres mots et expressions qui sont cooccurrents. Une fois interprétés, les cooccurrents pour lesquels on aura identifié une relation sémantique seront considérés comme des corrélats, c'est-à-dire comme des lexicalisations complémentaires de la même molécule sémique.
>
> Le réseau des corrélats relie les manifestations lexicales du thème. Mais il faut pouvoir discerner le(s) meilleur(s) point d'entrée(s) dans ce réseau. La « vedette » n'est alors qu'un mot d'entrée, choisi pour sa fréquence, et dans l'hypothèse qu'il présente une lexicalisation synthétique du thème que l'on cherche à décrire.
>
> (Rastier 1995a, §II.1, p. 236)

Les textes de l'ensemble des pages Web sélectionnées en réponse à une requête, soumise à un moteur de recherche sur Internet, est une forme de (sous-)corpus de ce type. Nos réserves viennent du fait que l'ensemble des pages indexées par le moteur est une réalité mouvante et mal définie, sans

---

[5] Pour une étude autour d'un personnage, voir par exemple (Dupuy 1993).

[6] Un des reproches adressés à une méthode d'Analyse du Discours, est de s'en tenir à un corpus de phrases extraites sélectionnées par des termes-pivot (mots pôles), sans y reconnaître un choix interprétatif majeur, qui relativise les résultats de l'étude à un point de vue :

« J.-M. Marandin, J. Guilhaumou et D. Maldidier, J.-J. Courtine, de manières convergentes, ont montré les limites d'une méthode *qui repose sur un savoir a priori*, celui qui préside à la sélection par le chercheur des termes-pivots : 'L'analyse répond à la question de l'analyste ; mais en présentant cette réponse comme structure de base d'un texte, l'analyste fait un passage à la limite où il confond son intérêt et ce qu'est le discours' [écrit J.-M. Marandin]. Choisir des termes-pivots, c'est définir les thèmes du discours ; or, dans la méthode des termes-pivots, ce n'est pas le texte qui permet de repérer ces thèmes, mais les présupposés de l'analyste » (Maingueneau 1991, §3.1, pp. 82-83)

Maingueneau poursuit en présentant une méthodologie de détermination de termes-pivots, qui s'appuie sur les constructions linguistiques :

« Courtine propose [...] de renverser ce problème de délimitation des thèmes du discours en posant la question suivante : '*Comment dans le discours lui-même et par le discours lui-même* un élément déterminé peut-il être caractérisé comme thème du discours ? (comment, c'est-à-dire : par la présence de quelles structures, sous quelle forme linguistique ?)'. [...] Cette option conduit naturellement Courtine à s'intéresser aux structures syntaxiques de la thématisation et, parmi celles-ci, tout particulièremet aux formules du type : 'C'est X que P', 'Ce que P c'est X', 'X c'est ce que P'. » (Maingueneau 1991, §3.1, p. 87)

logique d'ensemble, et donc ne forment pas un corpus qui vérifierait les critères énoncés ci-avant. Néanmoins, cet ensemble de pages, contenant toutes un ou plusieurs mots fixés, est un ensemble délimité, motivé, présentant une relative homogénéité thématique : il peut devenir le lieu d'analyses ciblées[7], inenvisageables sur Internet dans son ensemble.

Si la procédure de sélection des contextes est assez ciblée, la démarche peut devenir itérative : les nouveaux éléments contextuels sélectionnés lors d'une passe deviennent les pôles pour la passe suivante. C'est alors le fait que les contextes soient très sélectifs qui doit assurer la convergence, à savoir qu'après un certain nombre d'itérations, il n'y a plus d'éléments nouveaux sélectionnés (sans pour autant avoir sélectionné tous les mots du corpus !).

Greimas adopte une telle démarche itérative pour son étude à partir de l'opposition *vie* vs *mort*, dans l'œuvre de Bernanos :

> La procédure d'extraction apparaît donc, dans son ensemble, comme une série d'opérations successives d'extraction, chaque inventaire de contextes extraits permettant l'extraction et la mise en inventaire d'autres contextes, et ainsi jusqu'à épuisement du corpus, c'est-à-dire jusqu'au moment où la dernière extraction (*n*), utilisant le dernier inventaire (*n* - 1), ne fera plus apparaître de nouvelles qualifications. Cela voudra dire que le corpus utilisé pour fournir par extraction les éléments de signification appartenant à l'isotopie de *vie* et *mort*, choisie au départ, est épuisé de façon exhaustive. (Greimas 1966, §XII.1.b, p. 224)

Autre sélection motivée

Il s'agit de contraster un extrait (pas nécessairement continu), représentant une partie identifiée, par rapport au reste, ou encore de diviser (répartir) les textes en composantes homogènes. C'est ainsi que l'on peut choisir d'opposer les passages en style direct à la narration, les textes des différents personnages d'une pièce de théâtre.[8]

Echantillonnage

L'opération consiste à démultiplier un corpus pour pouvoir l'étudier sous la forme de plusieurs échantillons, chacun étant *a priori* représentatif de l'ensemble, avec quelques variantes locales jugées mineures (en fait, qui ne sont pas l'objet premier de l'analyse). L'avantage recherché peut être non seulement de posséder plusieurs « images » d'un même texte, mais aussi d'avoir des unités textuelles à caractériser pas trop longues, ou de taille régulière. En effet, la taille des échantillons est une décision extérieure au corpus.

> Un texte, considéré comme un ensemble, peut aussi être traité comme formé de plusieurs sous-ensembles. On peut soit en considérer les *divisions naturelles*, soit y créer des *divisions artificielles*. [...] Dans un roman, on pourrait considérer comme un sous-ensemble toutes les répliques des personnages (discours direct [...]), d'autre part tout le reste, où l'auteur se présente comme locuteur. Il va sans dire que ces divisions fournissent des fragments d'étendue variable et inégale. Même les cinq actes d'une pièce classique ont rarement le même nombre de vers. Nous réservons aux sous-ensembles ainsi créés [en suivant des divisions naturelles] le nom de *fragments*.
>
> On peut au contraire diviser un texte en *tranches* d'égale étendue sans tenir compte des divisions naturelles. On peut même constituer ces tranches par des segments prélevés en divers endroits du texte. Ainsi, pour diviser le texte de *Phèdre* (1 654 vers) en 10 tranches égales, à 1 vers près, donc de 165 ou 166 vers chacune, on peut soit couper aux vers 165, 330, 496, etc., soit prendre pour une première tranche les vers 1, 11, 21,..., 1641, 1651 ; pour une seconde les vers 2, 12, 22,..., 1642, 1652, et ainsi de suite [...]. Par ce dernier moyen, les tranches se rapprochent des échantillons aléatoires.
>
> [...] La réunion d'un grand nombre de textes indexés constitue un corpus, à l'intérieur duquel on se propose d'étudier et de quantifier certains faits lexicaux, syntaxiques, etc. Le corpus comprend en général des divisions ou sous-corpus, qui la plupart du temps ont une unité propre (chronologique, stylistique, etc.) ; dans ce cas ils entrent dans la catégorie des fragments telle qu'elle a été définie ci-

[7] Voir par exemple Live Topics, de François BOURDONCLE (Laboratoire de l'Ecole des Mines), qui opère sur les résultats de recherche d'Alta Vista.

[8] (Dupuy 1993, p. 261 sq.) prend la liberté de construire six « pseudo-textes », entre lesquels se répartissent les phrases de son corpus initial en fonction de leur « niveau dialogal » (est-elle prise en charge par le narrateur, par un personnage) et des temps de leur(s) verbe(s). La continuité linéaire des textes, et même les délimitations entre les différents textes du corpus (il s'agit d'un recueil de nouvelles) sont donc purement et simplement éliminées.

dessus ; s'ils ont été mesurés de façon à avoir la même étendue, ils se rapprochent des tranches, mais sans toutefois avoir été obtenus par une procédure aléatoire.

(Muller 1973, §3, pp. 15-16)

B. SÉMANTIQUE DES TEXTES, CALCULS ET MESURES

1. Des chiffres et des lettres : un mélange de genres ?

a) Croire aux chiffres : un outil puissant et suggestif

(Muller 1985) observe, chez une partie de ses collègues littéraires et linguistes, la crainte, l'hostilité, ou encore la fascination face aux statistiques, –trois attitudes également regrettables et auxquelles il invite à remédier.

Bien sûr, les statistiques sont par nature étrangères à la langue. Ceux qui restent sceptiques se trompent sans doute sur ses objectifs : il est effectivement inacceptable de penser que les statistiques rendent pleinement compte de la réalité linguistique. En revanche, c'est un mode d'investigation qui, méthodiquement employé, est en mesure d'apporter un soutien à l'analyste.

Certains reprocheront aux statistiques de ne s'intéresser qu'aux phénomènes de masse, sans s'intéresser aux singularités. Il est vrai que les statistiques visent à fournir une représentation globale et synthétique d'une réalité. Pour autant, un grand nombre de procédures statistiques servent à mettre en valeur ce qui s'écarte d'un comportement moyen ou uniforme. Par exemple, l'analyse factorielle met en valeur des pôles et des dimensions pour contraster les données, et dévalorise ce qui est proche de la moyenne ou intermédiaire. On recueille donc des informations sur des singularités relatives et significatives, « en contexte », plutôt que sur des singularités absolues et isolées.

> La statistique c'est l'art et la manière de réduire le support de l'information en perdant le moins possible d'information.
>
> La statistique c'est l'art et la manière de détecter les écarts aux grandes structures qu'elle a permis de mettre en évidence.
>
> La statistique c'est l'art et la manière de ne pas se laisser prendre dans la masse due aux effets de la combinatoire.
>
> La statistique c'est l'art et la manière d'aller à l'essentiel sans se perdre dans le détail.
>
> [...]
>
> La statistique devient donc l'art et la manière de poser des questions nouvelles à l'endroit qu'il convient dans le texte. Nous voulons donc, non pas résoudre des problèmes, mais les faire naître et même les faire naître par une lecture nouvelle, en indiquant où l'on doit relire. Le rôle du statisticien revient donc à dire : « allez lire à tel endroit ça ne fonctionne pas à cet endroit comme ailleurs, ça s'oppose à ce qui se passe à tel autre endroit ».
>
> (Massonie 1985, pp. 611-612)

b) Percevoir les limites du calcul pour mieux l'interpréter

Les sciences exactes sont (sur-)valorisées dans notre culture. Or il importe de garder un recul critique quant à l'emploi de la formalisation et des calculs. Tout calcul n'est pas « bon » en soi, que ce soit au sens de sa validité ou du respect d'une déontologie ; et toute terminaison de l'effectuation d'un algorithme n'est pas nécessairement un résultat scientifique. Il faut donc refuser de se laisser impressionner, et de céder à la fascination ou à l'aveuglement d'une mécanique brillante et complexe.

Il y a peu à gagner si l'on se focalise sur le côté calculatoire des traitements. Au lieu d'affiner la compréhension de ce qui est modélisé, et de repérer des leviers pour agir sur les données, on se perd dans une course au fignolage[9]. Premièrement, on ne peut attendre qu'un gain limité. L'expérience dans le domaine des calculs appliqués aux données textuelles enseigne que les résultats sont essentiellement dépendants de la pertinence et de la qualité des données, et que le choix de l'algorithme de classification (partitionnement) par exemple est beaucoup moins décisif que la constitution des données auxquelles l'appliquer (Quatrain & Beguinet 1996). Deuxièmement, on se condamne à une insatisfaction rémanente, car le seul but de la démarche est de s'approcher d'un idéal,

[9] « Si on analyse [...] les raisons profondes de ces premiers succès (Intelligence Artificielle combinatoire et Langage Naturel), on se rend compte qu'ils sont liés à l'utilisation de la *force brute* de la machine (en calcul et en mémoire) *utilisée intelligemment mais aussi simplement* » (Gondran 1994, Conclusion).

que l'on n'atteint jamais. Troisièmement, perdre de vue le lien à la réalité à décrire rend le traitement obtenu beaucoup plus vulnérable, car la réalité évolue.

De fait, l'utilisation de statistiques ne dégage pas de l'obligation de modéliser et de « penser » les données, bien au contraire. Et il ne faudrait pas opposer les traitements linguistiques et les traitements statistiques des langues en termes de compréhension des textes analysés. D'une part, les statistiques ne peuvent donner des résultats intéressants que si la nature linguistique et textuelle des données est prise en compte ; d'autre part, des outils de traitement morphologique, syntaxique ou sémantique, même les plus achevés, ne font jamais que préparer la compréhension, qui ne peut être faite que par un interprète humain.

Une approche trop naïvement quantitative se laisse détourner, et l'on assiste à des dérives comme celles du *spamming* pour les moteurs de recherche sur Internet. En effet, les moteurs étant généralement sensibles aux fréquences des mots, certains rédacteurs ont voulu abuser de cette propriété en commençant leur document par des dizaine de fois le même mot-clé répété, pour lui donner toutes les chances d'être mieux classé que les autres pages trouvées sur le même sujet, et d'avoir l'avantage d'être présenté dans les tout premiers. Des illustrations spectaculaires de ce procédé peuvent être trouvées dans (Koch 1996) -illustration 26 et suivantes ; le manuel de HOTBOT rapporte encore des ruses insoupçonnées :

> It has become popular for people to create pages that maliciously « spoof » search engines into returning pages that are irrelevant to the search at hand, or which rank higher that their relevance or content warrant. Common examples of spoofing are duplicating words thousands of times in comments or keywords, or icluding large number of « invisible » words in a tiny font, or in the same color as the background color of the page.

S'étant aperçu de la chose, certains moteurs de recherche ont alors tenté d'éliminer ces documents peu scrupuleux en rajoutant un seuil, avec une règle du genre : si un mot-clé apparaît plus de sept fois dans un document, alors celui-ci est présumé coupable de manœuvres. Mais ceci n'est évidemment pas infaillible...

L'opposé d'une approche crispée sur les formules et les chiffres, à savoir une utilisation trop qualitative des outils mathématiques, n'est pas bonne non plus. Chaque modélisation a ses lois, et les méconnaître fausse l'interprétation. En particulier, dans le cas des analyses factorielles, il n'est pas acceptable de s'en tenir à la première représentation plane obtenue, aussi suggestive soit-elle ; des indicateurs doivent être consultés et utilisés pour l'interprétation : pourcentage total d'inertie représentée, valeurs des contributions aux axes et des angles réels (avant projection), etc.

> *principe* : [il s'agit des] principes de l'*analyse* des *données* [...]. La sûreté d'une étude *statistique* requiert :
>
> - le *choix* des données sur une base délimitée sans conteste (*homogénéité*, *exhaustivité*) ;
> - un *codage*, traduction numérique (ou plutôt géométrique, dans un espace, muni d'une *distance* : cf. principe d'*équivalence* distributionnelle) fidèle au réel ;
> - un algorithme de *synthèse* par ordinateur conçu sans hypothèse *a priori* et éprouvé (cf. *comparaison*) ;
> - une *interprétation* centrée sur des *graphiques* dont la conformité aux données est rigoureusement mesurée (cf. principe *barycentrique*, principe du *bras* de levier, formule de *reconstitution*) et par lesquels l'étude de nombreux cas concrets suggère des *modèles* typiques reconnaissables *a posteriori*.
>
> (Benzécri & *al*. 1973, § *Indice systématique*)

> *interprétation* : en analyse multidimensionnelle, la validité est assurée par l'interprétation [qui s'appuie sur un certain nombre d'indices : *contributions*, éléments *supplémentaires*, etc.] (Benzécri & *al*. 1973, § *Indice systématique*)

c) Avertissement

Ce qui prime ici, dans la recherche de formules pour décrire les textes et les unités linguistiques qu'ils comportent, ce n'est pas l'exactitude d'un modèle, au sens où il est établi et démontré suivant un raisonnement formel. Ce qui importe avant tout, c'est l'adéquation des formules à la représentation des phénomènes textuels. Dans cette optique, les approches théoriques ou pragmatiques d'établissement des formules ont également droit de cité.

Pour les formules heuristiques, proposées comme expérimentalement efficaces, le but est de les expliquer : quelles sont les grandeurs qui interviennent, que représentent-elles, comment sont-elles combinées, quel est le comportement global saisi par cette formule (que met-elle en valeur), etc.

Pour les formules issues d'un modèle théorique, il convient d'expliciter les hypothèses sous-jacentes, notamment pour vérifier dans quelles mesure elles concordent avec la représentation que l'on se fait de la réalité textuelle et linguistique. C'est une bonne connaissance de la modélisation dont découle la formule qui guide ensuite l'interprétation des valeurs qu'elle fournit. Il est important également de prendre note des limites et conditions de validité prévues par la théorie. Une application de la formule hors du cadre prévu ne peut plus se référer au modèle théorique initial ; en revanche, elle peut encore être analysée d'un point de vue heuristique.

Certains modèles statistiques sont ainsi utilisés en interprétant l'écart entre le modèle et la réalité. Le modèle n'est pas une représentation parfaitement adaptée à la réalité textuelle, pour différentes raisons possibles. S'il s'agit d'un modèle de répartition des unités sur différents sous-corpus, les distorsions peuvent être :

- les sous-corpus correspondant aux divisions « naturelles » ou « logiques » du corpus (textes, collection d'extraits partageant une même caractéristique) ne sont pas des échantillons (représentatifs de l'ensemble du corpus), mais correspondent à des entités avec leurs particularités locales (la thématique propre à un texte, etc.)

  > si [...] le fragment a sa propre unité à l'intérieur [du corpus ou] de l'œuvre ([par exemple, dans *Le Cid*, les stances de Rodrigue, [ou] le récit de la bataille [...]), il est à prévoir qu'il aura ses caractéristiques propres. (Muller 1973, §3, p. 14)

- les unités ne se comportent pas comme dans un tirage aléatoire : la présence d'une unité dans un texte peut favoriser (ou au contraire inhiber) sa reprise dans le même texte. Autrement dit, chaque occurrence n'est pas indépendante des autres occurrences de la même unité.

  > [Expérimentalement,] la variance des effectifs [fréquences] observés est supérieure à celle des effectifs espérés [prédits par le modèle statistique]. [Autrement dit, pour une unité avec un nombre total d'occurrences fixé, et un ensemble de textes de longueurs connues, l'unité a tendance, dans chaque texte, a apparaître soit beaucoup, soit peu (ou pas).] [...] En réalité, la tendance du discours est bien la « spécialisation » qui entraîne une espèce de polarisation du vocabulaire dans les fragments [textes par exemple]. (Lafon 1980, p. 164)

  > Models for the distribution of words in text [...] are [often] inaccurate, as they ignore effects such as *word clustering*. Clustering arises from the fact that words tend to be repeated a number of times in the same piece of text, even words that are (overall) quite rare. (Thom & Zobel 1992, p. 616)

  (Lafon 1981a) met d'ailleurs en évidence l'apparition en « rafale » d'une proportion conséquente des mots dans un texte. Même s'il s'agit là de caractériser la répartition linéaire des mots à l'intérieur d'un texte, ce phénomène d'apparition localement groupée d'un mot peut également être un facteur d'irrégularité à l'échelle d'un corpus.

  > les « formes en rafales » sont massivement présentes dans les textes analysés, tandis que les formes régulières y sont rares. Selon les textes, 20 % à 35 % des formes retenues dans l'expérience ont une configuration exceptionnellement irrégulière [...]. Cette proportion élevée montre que le modèle utilisé (équiprobabilité des possibles), s'il permet de décrire, de juger, de classer avec efficacité, ne peut être considéré comme un modèle « représentatif » (simulant la réalité) des configurations observées dans un texte. C'est d'ailleurs, à notre connaissance, le cas de tous les modèles probabilistes utilisés en linguistique. Il faudra bien s'y habituer : l'emploi des mots dans un discours n'obéit pas, en général, à des lois simples de probabilité. La loi binomiale ou celle de Poisson, souvent utilisées dans les applications de statistique lexicale, ne peuvent être considérées, elles non plus, comme des modèles représentatifs, mais comme une autre approximation, plus ou moins commode, de celui que nous avons présenté. Car l'indépendance dans la succession des apparitions d'une forme, impliquée par ces lois, est, comme nous venons de le voir, trop souvent mise en défaut. (Lafon 1981a, pp. 186-187)

- les unités ne sont pas indépendantes : s'il s'agit de mots notamment, des interactions syntaxiques, sémantiques, stylistiques, favorisent l'apparition conjointe de certaines unités, et l'exclusion d'autres.

  > l'assimilation d'une partie du corpus à un échantillon du modèle est abusive. En effet, une partie est composée d'occurrences connexes qui se succèdent dans l'ordre naturel du discours. Cette propriété n'est pas respectée au sein des échantillons. Il est difficile de prévoir l'influence que cette distorsion peut avoir sur les résultats, il est vraisemblable qu'elle n'est pas uniforme et ne pèse pas

> toujours dans le même sens. Pour certaines formes [*i.e.* unités], il est possible qu'elle soit insignifiante. Ceci n'apporterait pas la preuve que ces formes ne tissent pas de liens contextuels (Lafon 1980, pp. 137-138)

Les valeurs de l'indicateur donné par le modèle « inadéquat » sont alors utilisables pour faire ressortir des points de divergence pour lesquels on propose une explication, mais pas pour avoir une représentation du corpus (qui permettrait par exemple des prédictions).

> Le modèle statistique, radicalement séparé du fonctionnement linguistique, est, en effet, tout à fait inapte à représenter celui-ci. C'est pourquoi nous ne lui demandons pas de nous fournir une approximation de la distribution des formes à travers les fragments d'un corpus. Il n'est pas imaginable d'employer le modèle [proposé, basé sur la loi hypergéométrique,] pour prévoir la composition d'un fragment ou pour prédire la sous-fréquence de telle ou telle forme [unité] dans une partie du corpus [...]. Le modèle statistique est de nature totalement étrangère à la réalité linguistique. Il n'est pas autre chose pour nous qu'un instrument de mesure permettant de détecter les formes qui justement s'éloignent le plus de lui, afin de donner une description précise de cette réalité. (Lafon 1980, p. 164)

2. La boîte à outils conceptuelle : un modèle général pour la comparaison des formules

a) Inventaire des grandeurs de base

Dans le domaine de la surveillance des installations (édifices, processus industriels), la question analogue est : où faut-il placer les capteurs ? Ce qui en fait recouvre des questions comme : de quelle nature sont les informations accessibles et enregistrables ? chacune peut-elle contribuer à cerner ce que l'on cherche à suivre ? et, réunies dans leur ensemble, sont-elles suffisantes ? autrement dit, quels sont les signaux indicatifs, utilisables, complémentaires ?

Où donc placer les « capteurs » de mesure sur les textes ? La réponse proposée ici consiste en une synthèse structurée, systématisée, des indicateurs utilisés. On en retire :

- un état des lieux,
- la mise en évidence d'une logique d'ensemble pour ces grandeurs,
- éventuellement, le signalement de virtualités peu ou pas exploitées,
- un jeu de notations unifié, utile pour des comparatifs.

Entrecroisement de deux espaces : paradigmatique et syntagmatique

Le corpus peut être mesuré selon deux dimensions : l'une, en termes de longueur, l'autre, en termes de diversité lexicale. Le tableau ci-après recense alors les différentes mesures envisageables, selon la combinatoire de ces deux dimensions :

	PARADIGMATIQUE			SYNTAGMATIQUE	
	Unités descriptives : unités reconnues et utilisées pour la représentation	**Unités élémentaires** : formes que prennent les unités descriptives dans les textes	**Occurrences** : réalisations des unités élémentaires	**Positions** discernables sur la linéarité	**Localisations** (typiquement les textes)
Total observable :	***U*** nombre d'unités descriptives, taille de l'univers	***N*** nombre de formes différentes relevées dans les textes, taille du vocabulaire	***F*** nombre de réalisations des unités	***L*** longueur	***T*** nombre de textes
Pour une unité *u* :	1	$\boldsymbol{N_u}$ nombre de formes différentes sous lesquelles se réalise *u*	$\boldsymbol{F_u}$ fréquence de *u* dans le corpus (nombre total d'apparitions)	$\boldsymbol{L_u}$ nombre de positions occupées par *u*	$\boldsymbol{T_u}$ nombre de textes dans lesquels *u* est présente
Pour un texte *t* :	$\boldsymbol{U_t}$ nombre d'unités descriptives caractérisant *t*	(N_t nombre de formes différentes dans *t*)	$\boldsymbol{F_t}$ nombre total d'occurrences dans *t*	$\boldsymbol{L_t}$ longueur du texte *t*	1
Pour une unité *u* et un texte *t* :	1	$\boldsymbol{N_{ut}}$ nombre de formes différentes sous lesquelles se réalise *u* dans *t*	$\boldsymbol{F_{ut}}$ fréquence (nombre d'occurrences) de l'unité *u* dans le texte *t*	$\boldsymbol{L_{ut}}$ portion de la longueur de *t* correspondant aux occurrences de *u* (présence)	1

Ce tableau peut surprendre par le nombre de variables qu'il identifie et distingue, et qui est notablement supérieur à ce que l'on observe dans les formules appliquées aux données textuelles.

En effet, la colonne des *unités élémentaires* (notations en *N*) est habituellement ignorée, et les modèles habituels décomptent directement les unités au fil du texte : soit l'unité est la forme graphique, soit des procédures de découpage et de réduction ont été appliquées (lemmatisation, reconnaissance d'expressions composées) et le rapport aux formes initiales n'est pas conservé.

Le tableau fait encore une autre distinction inhabituelle, entre les *occurrences* (en *F*) et les *positions* (en *L*) : en particulier, en une position peuvent apparaître plusieurs occurrences, et une occurrence peut s'étendre sur plusieurs positions. Cette idée n'est pourtant pas absolument nouvelle, c'est celle qui consiste à dissocier longueur et fréquences. Par exemple, (Muller 1977, §9, p. 51) indique explicitement, pour plusieurs formules, que telle variable est une mesure de la longueur du texte, et que pour celle-ci on peut adopter par exemple le nombre de vers, de lignes, ou de pages. Ce qui importe, c'est une indication des longueurs relatives des textes entre eux ; alors les mesures par le nombre d'occurrences, de vers, de lignes ou de pages sont équivalentes, car on passe de l'une à l'autre par une constante multiplicative (selon une approximation acceptable, et sachant qu'aucune n'est meilleure qu'une autre). En revanche, quand on a besoin de mesurer une proportion d'occurrences, alors c'est la variable F_{ut} qui est requise. En règle générale, les grandeurs relatives sont préférentiellement : par rapport à une unité F_{ut}/F_u, par rapport à un texte L_{ut}/L_t, ces grandeurs se confondant lorsque les occurrences se confondent avec les positions.

En ce qui concerne les localisations (notées en *T* et *t*), le tableau commente les variables correspondantes en se rapportant aux textes. Mais d'autres localisations syntagmatiques sont envisageables, et substituables au texte dans le tableau : le paragraphe, éventuellement la phrase ; ou encore le groupe de textes.

Dernière remarque : ce tableau ne fait aucune place à des mesures de distances syntagmatiques entre les occurrences (en nombre de positions par exemple). Il y a à cela plusieurs raisons. Le tableau tel qu'il se présente ici est suffisant pour décrire la quasi-totalité des formules relevées dans d'autres travaux et étudiées ci-après. La manière de définir cette grandeur est plus ouverte : ce peut être le nombre de positions entre la première et la dernière occurrences, ou le nombre moyen, minimum, ou maximum, de positions entre deux occurrences successives, etc. Dans le même mouvement, il faudrait ajouter des lignes pour définir des variables caractérisant un ensemble d'unités, un ensemble de textes, déclinant des mesures sur des unions et intersections (d'occurrences, de positions, etc.). On voit que le tableau en serait considérablement alourdi, et perdrait de sa clarté. Toutes ces grandeurs pourront néanmoins s'avérer utiles, et seront étudiées et nommées au moment nécessaire.

Propriétés possibles des segments

Les segments sont les entités définissables au plan syntagmatique, donc qui regroupent des occurrences (par opposition à des types). Nous nous intéressons en particulier aux entités vues aux chapitres précédents, et définies dans notre modèle des textes et du corpus. Il y a :

- des entités fondamentales : *corpus*, les *documents* qui présentent chacun un *texte*,
- des zones de localité plus fines : *paragraphes*, *périodes* (du même ordre que les phrases),
- des composantes liées à la structuration interne des textes : *parties*, regroupant des paragraphes en interrelation, par exemple formant une liste ; *surlignages*, mettant en valeur un paragraphe ou une suite de paragraphes, un mot (une unité élémentaire) ou une suite de mots ;
- des regroupements intertextuels : les *rangements*, qui définissent un ensemble de *boîtes*, dans lesquelles se répartissent les textes du corpus (un texte pouvant éventuellement appartenir à plusieurs boîtes ou à aucune) ; des *piles*, qui organisent les documents en *strates*, qui se superposent dans un certain ordre.

Les segments se décomposent jusqu'aux unités atomiques au plan syntagmatique : les occurrences des unités élémentaires.

Propriété 1 : Division

- *Définition* : segment défini parallèlement à d'autres segments (qui sont également des divisions), et tel que leur réunion soit équilibrée (les divisions sont toutes du même « niveau ») et représentative du corpus dans son ensemble. Les divisions peuvent se recouvrir, ou/et ne pas inclure une petite partie du corpus, qui ne serait pas pertinente pour le point de vue opéré par le découpage en divisions.
- *Illustration*
  exemple : un paragraphe (le corpus peut être décrit à travers l'ensemble de tous les paragraphes de tous les textes).
  contre-exemple : une partie (c'est un point du texte qui présente une organisation particulière, une partie n'entre *a priori* pas dans un système de parties qui couvrirait le texte) ; ou encore, la réunion des contextes d'un mot pôle (le « reste » du corpus ne constitue pas à proprement parler une entité de même nature que la sélection autour d'un mot pôle ; celle-ci est singulière. Cela pourrait devenir une division si l'on considérait par exemple toutes les réunions de contextes pour tous les mots du corpus, pris tour à tour comme mot pôle).
- *Signification opératoire* : l'étude de la répartition d'une unité sur l'ensemble des divisions d'un niveau fournit un indicateur de son comportement global sur le corpus.

Propriété 2 : Enclos

- *Définition* : segment qui vaut d'être considéré pour lui-même, qui présente une certaine autonomie, et qui reçoit ainsi une caractérisation.
- *Illustration*
  exemple : un texte.
  contre-exemple : un surlignage (il contribue seulement à modifier le rôle des unités concernées dans la représentation du texte).

- *Signification opératoire* : une caractérisation de l'enclos est calculée.

Propriété 3 : Progression

- *Définition* : segment qui institue une relation d'ordre entre ses composants directs.
- *Illustration*
  exemple : une phrase (ou période) (elle s'analyse en unités élémentaires, sur lesquelles est défini un ordre de succession).
  contre-exemple : une boîte (dedans, les textes sont « en vrac »).
- *Signification opératoire* : l'ordre permet de définir toutes sortes d'indicateurs supplémentaires :des proximités et distances, une orientation, des degrés d'intrication ou de recouvrement, etc. Si une progression est un enclos, alors son début et sa fin sont des points singuliers.

Bilan et application

Faisons le point des propriétés que présentent les segments que nous avons recensés (en italiques, les segments qui sont plus spécifiques à notre modèle) :

	division	enclos	progression
corpus		+	
pile			+
strate	+	+	
rangement			
boîte	+	+	
document	+	+	
texte		+	+
partie, rubrique, surligné			
paragraphe	+	+	+
période (phrase)	+	+	+

b) L'observable : que cherche-t-on à percevoir ?

Pondération intrinsèque (globale)

Il s'agit d'obtenir une qualification du comportement global d'une unité. C'est donc une fonction de l'unité *u*, mas pas d'un texte particulier *t* : P_u. Le plus souvent, on cherche une mesure de son pouvoir de discrimination, à savoir de sa capacité à différencier les textes du corpus. Les indicateurs font appel à des notions comme la dispersion, l'entropie.

Pondération contextuelle (locale)

Le schéma classique de pondération d'un mot-clé, dans les systèmes documentaires comme le SMART de l'équipe de Salton, est le produit de trois valeurs (Hersh & *al.* 1994) :

- une mesure de la réalisation du mot dans le document ; par exemple :
  - une fonction binaire, qui vaut 1 si le mot est présent,, 0 sinon ;
  - le nombre d'occurrences du mot dans le document ;
  - une fonction croissante du nombre d'occurrences, qui a de plus certaines propriétés particulières : atténuation des hautes fréquences, caractère borné, etc.
- une mesure du caractère discriminant du mot dans le corpus, typiquement l'*inverse document frequency* (*idf*), par exemple de la forme $\log(T/T_u)$;
- un facteur de normalisation, qui égalise les poids totaux de tous les documents ; par exemple, la division par la norme (au sens du cosinus) du vecteur pondéré représentant le document.

La pondération contextuelle reprend donc habituellement les informations de la pondération intrinsèque (la deuxième des trois valeurs ci-dessus), et éventuellement prépare le calcul de similarité en réajustant la pondération résultante pour l'ensemble du texte.

La pondération locale évalue l'importance de l'unité *u* pour le texte *t* : c'est une fonction de *u* et *t*, P_{ut}.

Similarité

Le calcul peut aussi chercher à évaluer la force du lien entre deux textes : grosso modo, y a-t-il une pertinence à considérer ces documents ensemble, l'un appelle-t-il l'autre, ou bien sont-ils sans rapport significatif ?

Dans les moteurs de recherche actuels sur le texte intégral (type recherche d'informations sur Internet (Pincemin, Lemesle 1996, §2.III.B) ou moteurs généraux commercialisés tels que PLS – *Personal Library Software* (Banet 1996)), la similarité entre une série de mots-clés (la requête) et un document (le contenu de la page Web) est une fonction croissante d'indicateurs parmi ceux-ci :

- la rareté de chaque mot-clé et son caractère discriminant (le mot-clé apparaît dans peu de documents, dans l'ensemble des documents recensés)
- la fréquence du mot-clé dans le document, qui se décline en fréquence absolue (le mot-clé a un grand nombre d'occurrences dans le document), fréquence relative (il a un grand nombre d'occurrences par raport à la longueur totale du document), fréquence suffisante (par exemple, ce n'est pas un hapax, il apparaît plusieurs fois dans le document).
- la position de chaque mot-clé dans le document, si elle est remarquable : le mot-clé est-il proche du début du texte, ou est-il dans une zone significative (il fait partie du titre, des mots-clés indiqués par le rédacteur de la page).
- la présence du mot-clé sous sa forme exacte, plutôt que dérivée : le mot-clé est retrouvé dans le document sous une forme identique à celle qu'il a dans la requête, et non sous une de ses variantes.
- le nombre de mots-clés de la requête présents dans le document (ce critère étant d'autant plus important que la requête est brève, et que seule la présence d'un certain nombre de mots-clés assure une contextualisation minimale et désambiguïsante).
- la proximité des mots-clés de la requête dans le document : il s'agit d'éviter les cas où la présence simultanée de plusieurs mots-clés dans le document est fortuite, en repérant des facteurs pouvant refléter leur interrelation effective dans le document (la présence dans un même passage pour une affinité sémantique, la succession immédiate pour reconstituer un terme composé ou une expression).
- l'ordre de deux mots-clés est le même dans la requête et dans le document : cet indicateur vise à tenir compte d'expressions composées, qui ont été désarticulées par l'atomisation de la requête en mots-clés simples.

On reconnaît, dans les premiers indicateurs, les formules de pondérations classiques. Les deux indicateurs suivants sont des modes complémentaires et plus élaborés de pondération contextuelle (il s'agit toujours de la relation entre un mot-clé et un document). Avec les trois derniers points, entrent en ligne de compte les interactions de plusieurs mots-clés dans le texte. Ces trois points corrigent tant bien que mal les effets négatifs du morcellement initial de la requête en mots-clés indépendants. Ils supposent aussi implicitement que l'on a affaire à une requête de quelques mots-clés, et ne s'étendent pas tels quels au calcul de similarité entre deux textes intégraux.

La fonction de similarité s'applique à une paire de textes. Peut-être une mesure d'homogénéité, applicable à un ensemble de textes, pourrait être une généralisation d'une fonction de similarité.

L'indicateur numérique (proximité ou distance) vise à informer sur un degré d'équivalence, mais ne caractérise pas directement le mode de cette équivalence (à savoir : sous quels rapports tel et tel textes sont-ils proches). La linguistique s'est posé le problème de l'équivalence sémantique entre mots (synonymie), phrases (paraphrase) ; il reste à poser le problème herméneutique de l'équivalence entre textes.

> à quelles conditions un texte en représente-t-il un autre, et à quelle condition cette représentation apparaît-elle équivalente ? (Rastier, Cavazza, Abeillé 1994, Epilogue §2.4, p. 206)

Interprétation relative ou absolue

Les modèles statistiques sont habituellement utilisés pour tester des hypothèses : par exemple ici, « telle unité est uniformément répartie et peu caractérisante (elle doit recevoir un poids faible) ». Cela suppose de fixer des seuils : à partir de telle valeur de la mesure statistique, on peut considérer

que l'hypothèse d'équirépartition n'est pas vérifiée, c'est donc que l'on a affaire à une unité significativement caractérisante.

Une autre manière de procéder, qui évite de devoir trancher en fixant un seuil, consiste à garder toutes les valeurs de la mesure statistique et à les considérer par rapport à l'ordre qu'elles instaurent (Muller 1973, §17, pp. 114-115) (Muller 1977, §9, pp. 52-53). Ainsi, on a des informations telles que : telle unité est plus discriminante que telle autre ; les *n* unités les plus spécifiques d'un texte ou les plus irrégulières dans le corpus, et de façon duale, les *n* unités les plus anormalement peu fréquentes ou absentes.

3. La mise en formule : choix et signification - Exemples d'analyse

a) Fréquences

Intuition de départ

Il y a manifestement un certain réalisme des fréquences : la fréquence traduit la reprise, et effectivement le locuteur ou le scripteur ne réinvente pas le choix et les combinaisons des mots à tout moment. Le dictionnaire s'appuie sur le déjà-dit, à travers les exemples donnés (qui y gagnent même une certaine canonicité).

> La description sémantique considère la répétition, et, par là même, la fréquence relative des éléments itératifs du contenu, comme un phénomène normal, et non comme investie d'un statut particulier. La fréquence, dans un texte donné, d'éléments à formants identiques est un indice utile, révélateur de redondances camouflées probables, et son rôle, sur le plan pratique, n'est pas négligeable. (Greimas 1966, §IX.3.b, p. 160)

Observation : loi de Zipf

La loi de Zipf est une formule heuristique, remarquable en ce que l'on observe qu'elle s'applique *toujours*, quel que soit le texte considéré, avec une approximation minime. Elle refléterait donc une propriété fondamentale du système sémiotique de la langue.

La loi de Zipf énonce que, si l'on ordonne les mots d'un texte par fréquence décroissante et si on leur attribue un rang, alors le rang trouvé est à peu près inversement proportionnel à la fréquence. L'exemple classique est l'*Ulysses* de Joyce :

- le mot de rang 10 (*i.e.* il n'y a que 9 mots plus fréquents que lui) a une fréquence de 2 653 ;
- le mot de rang 100 (*i.e.* il n'y a que 99 mots plus fréquents que lui) a une fréquence de 265 ;
- le mot de rang 1 000 (*i.e.* il n'y a que 999 mots plus fréquents que lui) a une fréquence de 26 ;
- le mot de rang 10 000 (*i.e.* il n'y a que 9 999 mots plus fréquents que lui) a une fréquence de 2.

La représentation des paires *rang - fréquence* en coordonnées log-log (diagramme de Pareto) permet d'observer immédiatement la loi, puisqu'alors l'ensemble des points se dispose théoriquement selon une droite de pente -1.

Cette loi peut-elle être exploitée pour la caractérisation des textes ? Peut-elle apporter une information d'ordre sémantique ? Caractériser, c'est notamment distinguer, positionner dans un environnement. Or la loi prédit une répétition à l'identique, un comportement uniforme des lexiques des textes. Ce sont donc les écarts à la loi, observés pour tel ou tel texte, qui donneraient matière à observation et interprétation (Bommier 1993, §I.d, pp. 9-10), sachant que les spécialistes s'accordent sur le fait que la loi est communément moins bien vérifiée aux valeurs extrêmes (Guilbaud 1980) (Mandelbrot 1968).

L'évaluation automatique de ces écarts reste problématique, et le gain sémantique à en attendre très incertain : mieux vaut garder cette loi comme une curiosité intéressante, à laquelle confronter les corpus rencontrés, pour avoir une image de leur « trace sémiotique »... Toujours est-il que l'existence de ce phénomène sur la distribution des fréquences convaincra que les fréquences ne prennent pas « n'importe quelles » valeurs, et qu'il y a là un certain ordre du texte.

Hapax

Est communément appelé *hapax* un mot de fréquence 1.

Point de discontinuité

La fréquence 1 est un point sensible car doublement remarquable : c'est *(i)* ce qui est présent (fréquence supérieure à 0), et *(ii)* ce qui n'est pas répété (fréquence inférieure à 2). Or ces deux observations conduisent à des orientations tout à fait différentes, ce qui fait de cette fréquence 1 une valeur éminemment sensible.

L'unité de fréquence 1 est présente dans le texte : elle fait donc partie du relatif petit nombre des unités relevées dans le texte, par rapport à l'ensemble des unités connues par le corpus de référence. Elle porte donc la marque d'un « choix », celui qu'a fait l'auteur du texte de l'utiliser. (Cela ne présume pas du caractère original et sélectif de ce choix, qui se mesure par un indicateur inter-textuel, de spécificité).

L'unité n'est pas répétée : dans la pratique, on observe que les hapax drainent l'essentiel des éventuelles fautes de frappe, des termes aberrants, exceptionnels, peu significatifs pour le corpus. Qualitativement, ils concentrent des termes instables et porteurs de bruit (c'est-à-dire favorisant les erreurs dans le traitement automatique). La répétition fonctionne souvent comme une confirmation, qui fait donc défaut au hapax.

Autre caractère paradoxal des hapax, ce sont eux les plus présents dans l'inventaire lexical d'un corpus, et en même temps il est délicat de leur donner un rôle dans le traitement automatique, par manque d'information sur leurs affinités contextuelles *stables*. Leur importance quantitative est indéniable : quelles que soient les textes considérés, près de la moitié des mots sont des hapax (c'est un ordre de grandeur ; voir par exemple (Guilbaud 1980) et ses exemples illustratifs). Mais ils sont évidemment inertes dès lors qu'il s'agit d'établir des termes de comparaison internes au corpus, ou pour repérer des régularités.

Hapax de mot et hapax de texte

En adaptant les propositions de (Bommier, Lemesle 1995), définissons le *hapax de mot* (mot qui n'apparaît qu'une fois dans le texte considéré) et le *hapax de texte* (mot qui n'apparaît que dans un seul texte du corpus).

Il y aurait alors trois formes de réalisation d'une fréquence 1, dont la prise en compte permet des interprétations plus précises.

Pour une unité qui est à la fois hapax de mot et hapax de texte :

- comportement typique d'une faute de frappe

Pour une unité qui est hapax de texte seulement (elle apparaît plusieurs fois dans le texte, mais pas en-dehors) :

- mot spécifique, mais qui peut l'être à divers titres (notamment selon les corpus) : erreur sur l'orthographe d'usage ; terme pointu lié à un domaine d'expertise ; néologisme ou nom propre (lié à une personne ou à une situation) ;
- utilité incertaine, mais potentiellement déterminante, pour l'étude de liens avec des textes hors du corpus.

Pour une unité qui est hapax de mot seulement (elle n'apparaît qu'une seule fois dans le texte, mais elle apparaît également dans d'autres textes) :

- ne joue peut-être jamais un rôle thématique majeur. Cette hypothèse est renforcée par la longueur des textes, et l'uniformité de sa distribution sur les textes du corpus.

La distinction entre hapax de mot et hapax de texte ne joue pleinement que pour des corpus considérant des textes suffisamment longs (par exemple, pour la presse, des articles développés par opposition à des dépêches). Pour des textes de type résumé, il peut être rare que les unités lexicales soient répétées au sein d'un résumé. Sur une page Web, le procédé des liens hypertextes permet de ne mentionner qu'une seule fois une notion importante, qui est développée dans une page attenante.

> Pour les documents EDF, les résumés s'apparentent à une énumération de sujets qui ne sont généralement pas répétés dans le même résumé, d'où la nécessité d'une statistique globale pour repérer les termes composés significatifs de l'ensemble des documents. Dans le cas des brevets DERWENT, au contraire, la répétition locale des termes est notable pour les résumés. Si l'on veut prendre en compte une spécificité et une exhaustivité optimale concernant les termes à extraire, il est alors intéressant de doubler les statistiques globales par des statistiques locales au document. (Chartron 1988, §VIII.2, p. 111)

> Hypertext, however, has a property which allows people to mention topical words only once without losing the cohesion of the document. This means that when trying to identify clumps of identical topical words within a hpertext document, the behaviour of the word frequency might be diffferent. Since lists and indexes are topical words organised in a methodical way, and very often are not repeated in the content of the document, there is a need to test the assumptions made for text, on hypertext. Topical words might reoccur in related or linked documents, but if the documents are collected randomly and are treated as once-occurring HTML files, then the word frequency behaviour should be expected to be different from that of words in flat texts. (Amitay 1997, §2, p. 19)

De même, typiquement, les requêtes documentaires par mots-clés ne comportent quasiment que des hapax de texte, au sens où, conventionnellement, dans la requête, chaque mot-clé n'est donné qu'une fois. Dans la cas où l'on a à traiter ce type de « texte », il ne faut donc surtout pas éliminer les hapax de document !

- *tf.idf (term-frequency × inverse document-frequency)*

Appliquons les observations précédentes à la très classique pondération *tf.idf* de Salton, à savoir « term-frequency × inverse document-frequency », c'est-à-dire :

$$P_{ut} = F_{ut} \times \log\left(\frac{T}{T_u}\right)$$

ou sa version employée dans ADOC (moteur de la version 1 de DECID) :

$$P_{ut} = F_{ut} \times \left(1 + \log\left(\frac{T}{T_u}\right)\right)$$

Un mot à propos de cette variante : la constante additive permet d'atténuer l'incidence trop directe de l'*inverse document-frequency* (*idf*) et notamment de ne pas dévaluer trop vigoureusement et irréversiblement les unités présentes dans une majorité de documents. Elle est fixée ici relativement arbitrairement à 1 : l'effet correctif et atténuateur peut être modulé en changeant sa valeur (plus la constante est grande, moins l'influence de l'*idf* se fait sentir.

La formule s'attire des remarques, par rapport au traitement des hapax :

- Les hapax de textes sont toujours valorisés, y compris ceux qui sont également hapax de mot.
  Une première manière de corriger cela, est de remplacer le facteur F_{ut} par une fonction $f(F_{ut})$, qui *(i)* dévalorise la fréquence 1 (source importante de perturbations), *(ii)* revalorise la fréquence 2 (pour accentuer la différence d'avec les hapax, la répétition étant un indice de fiabilité), *(iii)* valorise encore la fréquence 3 (qui apporte une confirmation plus marquée que la fréquence 2), *(iv)* reste une fonction croissante de la fréquence, mais avec des différences de moins en moins sensibles à mesure que la fréquence croît (on retient surtout que l'unité est répétée plus de 3 fois). Par exemple :
  $f(F_{ut}=1) = 0.4$; $f(F_{ut}=2) = 2$; $f(F_{ut}=3) = 5$; $f(F_{ut}) = 5+\ln(F_{ut}-2)$ pour $F_{ut}>3$.
- La valorisation est en fait très modérée. Le logarithme bride les valeurs fortes : le logarithme est majoré par le logarithme du nombre total de documents dans le corpus ; mais, surtout si les mots grammaticaux ont été éliminés, le logarithme ne prend que très rarement une valeur faible. La fourchette de variation du logarithme est réduite.
- La variation de la pondération est d'autant plus grande que le terme apparaît dans peu de documents, et la pondération aura une valeur sensiblement différente si le mot apparaît dans 1, ou 2, ou 3 documents par exemple.
- Si ensuite on utilise une mesure de similarité entre vecteurs normalisés, le choix de cette pondération fait que les textes comprenant plus de hapax de texte que la moyenne sont défavorisés et n'obtiennent que des similarités faibles. Autrement dit, un texte qui a des développements originaux peine à être repéré, même s'il a par ailleurs d'autres parties plus classiques qui lui donnent des points communs avec d'autres textes.

Influence du type d'indexation

L'allure des fréquences n'est pas la même selon que l'on caractérise les textes par les mots qu'on y trouve (approche type « texte intégral » et fichier inverse), ou que l'on opère une indexation contrôlée (par une terminologie, un thesaurus) tout en gardant une indication du nombre de fois où un descripteur a été relevé dans le texte.

En effet, dans l'approche type « texte intégral » et fichier inverse, rien ne prémunit la caractérisation contre les fautes de frappes et les spécificités lexicales d'un rédacteur (idiolecte) : celles-ci sont les principales sources qui alimentent les hapax de mot et hapax de texte.

En revanche, le passage par un vocabulaire contrôlé filtre les particularismes et les anomalies (qui font classiquement l'essentiel des termes non reconnus). Les hapax n'ont donc plus la même nature que dans le cas précédent, ils ne peuvent plus être soupçonnés du fait de leur non attestation, car ils sont par définition enregistrés, donc validés, dans le vocabulaire contrôlé. En revanche la proportion de hapax ne varie pas significativement (pour notre expérience sur les textes d'Action), le fait de travailler en vocabulaire fermé n'engendre pas nécessairement une accumulation (fortes fréquences) sur les descripteurs trouvés.

Fréquence et répétition

La fréquence est une mesure atténuée de la répétition sémantique. Cette affirmation appelle quelques explications, que nous précédons d'une remarque immédiate : il convient de reconnaître encore une fois que l'on n'a pas de mesure « directe » du sens à partir des formes observables.

Par ailleurs, la langue a « inventé » une manière légère, quasi imperceptible, de poursuivre le fil de son discours sans reprendre, continuellement et explicitement, les désignations développées des objets déjà introduits : on aura reconnu le mécanisme des pronoms, et plus généralement celui des anaphores.

De plus, les principes habituels de rédaction, dans notre culture, prônent d'éviter les répétitions : et l'on apprend à l'écolier à s'ingénier à varier son vocabulaire pour satisfaire à cette esthétique.

Ainsi, une notion sémantique apparaît de façon diffuse dans le texte : dans les relations contextuelles entre les mots, dans des expressions variées, et même dans des unités quasi neutres et « passe partout ». Vouloir déterminer chaque manifestation d'une notion sémantique pour la comptabiliser de façon exacte est cependant une voie illusoire. Premièrement, l'automatisation du repérage des anaphores ou des reprises lexicales est extrêmement complexe et non résolue. Deuxièmement, un tel décompte n'est qu'une autre traduction de l'effet sémantique, pas plus exacte que les fréquences (la réalité sémantique n'est pas de nature comptable) ; et un bon usage des fréquences est *a priori* tout aussi efficace que le serait cet indicateur de fréquences « réévaluées ». De fait, cet indicateur ou les fréquences ont un comportement global analogue : une idée constamment reprise dans un texte se traduit par une répétition, notamment de paragraphe en paragraphe, que captent les fréquences. En ce sens, les fréquences ont le même comportement que cet indicateur imaginé, pour des valeurs numériquement simplement plus faibles. D'où l'adjectif « atténué » dans la proposition initiale : « la fréquence est une mesure atténuée de la répétition sémantique ».

A l'inverse, on peut à juste titre objecter que l'identité des occurrences n'implique pas l'identité des types, et qu'un signifiant, en fonction de son contexte, peut renvoyer à des signifiés très différents.

> Que vaut l'addition lorsque les éléments sont disparates, mêmes s'ils revêtent le même habit ? Quand on cumule l'*amour* du jeu, l'*amour* du prochain et tant d'autres *amours*, licites ou interdits, procède-t-on à l'intersection des sèmes communs, ou bien s'agit-il de la réunion de sèmes divergents ? Le même effectif peut dans le premier cas recouvrir un sous-ensemble vide, dans le second un sur-ensemble débordant. Malgré leur précision, les chiffres obtenus sont-ils autre chose que des ballons gonflables, sensibles à la pression alternée de l'extension et de l'intension ? Nul n'a jamais répondu définitivement à cette objection majeure –qui n'est pas propre au langage et qui obère pareillement l'ensemble des sciences humaines et même la plupart des sciences d'observation. Avant de conclure à la répétition d'un phénomène –ce qui permet de conclure à quelque loi–, comment être sûr que les conditions sont rigoureusement semblables et qu'il s'agit bien de deux occurrences d'un même événement plutôt que de la coïncidence de deux séries causales ? (Brunet 1995, p. 26)

Les effets de polysémie ou d'homonymie sont cependant limités dans le cadre d'un texte scientifique et technique, les collisions de signifiés étant rares dans un domaine délimité par une pratique. D'autre part, la réitération d'un signifiant appelle une mise en relation interprétative des occurrences.

> Points fixes placés sur la ligne mouvante du discours, [les séquences de mots itérées] instaurent, comme les rimes en poésie, une équivalence relative entre les phrases qui les portent : d'où des effets de sens variés, résultant des oppositions ou des identifications qui s'opèrent. (Dupuy 1993, p. 119)

Finalement deux occurrences n'ont ni exactement le même sens, ni deux sens totalement étrangers l'un à l'autre, et il est illusoire de croire tenir la vérité en optant pour l'une ou l'autre de ces deux solutions, –confondre totalement ou séparer totalement. C'est pourtant ce que nous obligent à faire les calculs, et cela n'est acceptable qu'en y reconnaissant une approximation transitoire. La morale de tout cela, est encore une fois qu'il faut se garder de croire décompter avec exactitude des éléments de sens : la nature profonde de la sémantique n'est pas de l'ordre d'un calcul, exact et déterministe.

Limites de validité et fausses intuitions

Certaines formules sont établies en fonction de modèles statistiques, et leur domaine de validité ne couvre pas toute la gamme des fréquences.

- *Ecart réduit et* χ^2

Pour une unité donnée, l'écart réduit sert à comparer sa fréquence réalisée dans un texte particulier par rapport à sa fréquence générale dans un corpus étendu (Muller 1977, §9, pp. 49-50). Plus l'écart réduit est proche de zéro, plus le texte se présente comme un échantillon représentatif du corpus :

$$Ecart\_Réduit_{ut} = \frac{Fréquence\_Observée - Fréquence\_Théorique}{\sqrt{Fréquence\_Théorique}} = \frac{F_{ut} - \left(F_u \times \frac{L_t}{L}\right)}{\sqrt{F_u \times \frac{L_t}{L}}}$$

Remarque : Avec les grandeurs que nous considérons, le sous-corpus pour lequel est calculé un écart réduit est un texte, et le corpus d'établissement des fréquences théoriques est l'ensemble des textes étudiés. L'écart réduit peut bien sûr être mis à profit pour décrire d'autres types de sous-corpus, et par rapport à un autre corpus de référence..

Le χ^2 sert à apprécier en probabilité l'écart constaté entre une observation et un modèle théorique (Muller 1973, §18). Il est d'application plus générale que l'écart réduit car il est capable de prendre en compte l'écart sur plusieurs variables (sa valeur s'interprète d'ailleurs en fonction du *nombre de degrés de liberté* sur l'ensemble des variables considérées) :

$$\chi^2(u) = \sum_{t=1}^{T} \frac{\left(Fréquence\_Observée - Fréquence\_Théorique\right)^2}{Fréquence\_Théorique} = \sum_{t=1}^{T} \frac{\left[F_{ut} - \left(F_u \times \frac{L_t}{L}\right)\right]^2}{F_u \times \frac{L_t}{L}}$$

Remarques : la formule ci-dessus est une application particulière du test du χ^2 à la répartition d'une unité parmi des textes.

Le nombre de degré de liberté vaut (T-1), il est constant quelle que soit l'unité u. On peut donc comparer entre elles les valeurs de $\chi^2(u)$, pour différents u.

Une formulation théoriquement plus juste prendrait en compte non seulement les écarts de l'unité considérée, mais aussi les écarts qui en découlent pour les autres unités. L'approximation est cependant acceptable, car la variation occasionnée par une unité sur la répartition de toutes les autres dans leur ensemble est négligeable devant la variation pour l'unité elle-même.

> Pour le χ^2, on pourrait faire entrer dans le calcul non seulement les affectifs du groupe de vocables considéré, mais le total des autres mots, qui lui est complémentaire [...]. Le résultat n'en serait pas affecté sensiblement, parce que la fréquence des vocables considérés, par rapport au nombre

> total d'occurrences du texte, est très faible (inférieure, avons-nous admis, à 0,01, et très souvent à 0,001). La prise en compte des autres mots serait donc sans intérêt, sauf si l'on opérait sur des unités ou des groupes à fréquence plus élevée, ce qui est tout à fait exceptionnel en statistique lexicale. (Muller 1977, §9, pp. 50-51)

La présentation de ces formules s'accompagne nécessairement des indications sur leurs limites d'emploi : en particulier, elles ne sont pas adaptées pour des unités rares par rapport aux divisions considérées. Les formules pourraient donc mieux convenir pour caractériser des regroupements de textes ou des unités assez génériques. Si l'on étend leur utilisation aux domaines de basse fréquence, les résultats ne sont plus exploitables en termes de probabilité (test d'une hypothèse), mais peuvent néanmoins donner des mesures relatives pour comparer les unités entre elles.

> 1. [Le test du χ^2] ne s'applique qu'à des *effectifs* [*i.e.* fréquences] *absolus*, jamais à des effectifs relatifs (pourcentages, par exemple), et jamais à des grandeurs (caractères quantitatifs) ;
>
> 2. il perd de sa précision quand les écarts sont établis par rapport à des *effectifs théoriques trop faibles* ; dans la pratique, on s'interdira d'inscrire dans les effectifs théoriques des nombres *inférieurs à 5*, et on évitera s'il se peut les nombres inférieurs à 10 ; pour les effectifs réels, il n'y a pas de limitation ;
>
> 3. toute erreur sur le nombre de degrés de liberté fausse le résultat.
>
> [...] Ces trois prescriptions sont d'importance très inégale.
>
> La limite de 5 est arbitraire ; certains préfèrent la placer à 10 ; d'autres s'en affranchissent volontiers, surtout quand le but du test est moins d'apprécier un écart en probabilité que de comparer des distributions entre elles. Quant aux degrés de liberté, il y a des cas où leur nombre prête à contestation.
>
> En revanche l'obligation de n'opérer que sur des effectifs réels est absolue.
>
> (Muller 1973, §18, pp. 121-122)

Quant aux fortes fréquences, elles font pleinement partie du domaine de l'écart réduit, mais auraient tendance à être valorisées, ce qui conduit à une inversion des effets habituellement observés lorsqu'elle sert de point d'entrée à une analyse factorielle :

> Comme l'analyse a été faite sur les écarts réduits l'effet de taille ne joue pas –qui habituellement ramène vers le centre de gravité les effectifs les plus lourds. On constate plutôt l'effet inverse qui est d'éloigner du centre [situé au croisement des axes de l'analyse factorielle] les écarts les plus forts. Car les écarts réduits ont tendance à augmenter quand agit la loi des grands nombres. La perspective se trouve inversée : au lieu de constituer la norme indistincte, près de l'origine des axes, les gros effectifs deviennent discriminants et attirent l'attention sur les marges qu'ils atteignent plus rapidement, comme ces nébuleuses qui fuient aux confins de l'univers et dont la vitesse est proportionnelle à leur masse. (Brunet 1995, note 16, p. 52)

La loi de Poisson est au contraire adaptée à la description des basses fréquences (mais peu appropriée pour des fréquences moyennes ou élevées).

> Quand le calcul de la fréquence théorique [*i.e.* ($F_u \times L_t / L$)] porte sur des fréquences assez faibles et quand le corpus de référence est très grand par rapport au texte, il arrive que la fréquence théorique soit peu supérieure ou même inférieure à 1 ; dans ce cas (en pratique en-dessous de 5), il est peu indiqué de recourir au χ^2 et à l'écart réduit ; on préférera appliquer la loi de Poisson. (Muller 1977, §9, p. 53)

(Lafon 1980) modélise la répartition des unités sur les textes d'un corpus comme une distribution hypergéométrique : il obtient ainsi un indicateur valide sur toute la gamme des fréquences. Il observe alors plusieurs particularités, contre-intuitives et importantes, pour les unités de forte fréquence.

Ses expérimentations montrent en particulier qu'un raisonnement sur les proportions relatives des occurrences d'une unité dans différents textes (*i.e.* basé sur la quantité F_{ut}/F_u) introduit des équivalences trompeuses. En effet, il est statistiquement beaucoup plus significatif, pour une unité très fréquente que pour une unité rare, de s'écarter d'une distribution donnée.

> Notons 220/801 l'événement suivant : « Une forme [unité] répétée 801 fois dans le corpus, figure au moins 220 fois dans l'échantillon ». Le calcul nous montre [par exemple] que les événements 220/801, 112/365, 45/125, 29/71, 23/52, 8/11, 6/7, 5/5, sont équivalents en ce sens que leurs probabilités sont sensiblement égales. Ont-ils pour cela la même importance ? Nous ne nous prononcerons pas ici sur cette délicate question. Ce que nous voulons faire sentir c'est que l'évolution

> des rapports qui traduisent des événements équivalents n'est pas linéaire, contrairement à l'idée qui vient spontanément à l'esprit. On voit que les [rapports fréquence locale / fréquence totale] sont loin d'être proportionnels, à l'inverse de ce qu'une intuition trompeuse et largement répandue invite trop souvent à penser. (Lafon 1980, p. 159)

De même, les fréquences relatives, à savoir rapportées à la taille des textes, (il s'agit dans notre notation de L_{ut}/L_t) mésestiment les fortes fréquences : ce point sera repris à propos de la prise en compte de la taille des textes (car la fréquence relative serait une manière intuitive de normer les fréquences).

D'une manière générale, les unités très fréquentes se caractérisent plutôt par leur irrégularité que par leur distribution uniforme :

> Ainsi, la répétition massive d'une forme [*i.e.* une unité], qu'elle soit fonctionnelle ou lexicale, se conjugue rarement avec une distribution régulière. (Lafon 1980, p. 158)

Calibrage : la fréquence maximale

Autant les fréquences minimales sont connues d'avance (il y a toujours des hapax, en nombre), autant les fréquences maximales ne sont pas aussi aisément prévisibles. Les majorants évidents sont très éloignés des majorants réels. Ceux-ci sont déterminables pour le corpus de référence : ils donnent un ordre de grandeur plus adapté pour des facteurs de normalisation. En particulier, la fréquence de l'unité la plus fréquente dans un texte ne varie pas linéairement avec la longueur du texte : relativiser les fréquences par rapport à la fréquence maximale dans le texte est plus juste que de normaliser par la longueur du texte.

C'est ainsi que l'on peut comprendre la formule de pondération suivante, dans la lignée des formules saltoniennes (Hersh & *al.* 1994) :

$$P_{ut} = \left[0{,}5 + 0{,}5 \times \left(\frac{F_{ut}}{\max\limits_{u \in Vocabulaire} \left(F_{ut} \right)} \right) \right] \times \sqrt{\log_2 \left(\frac{T}{T_u} \right)}$$

Présence / absence

L'indication de présence / absence est en soi moins informative que l'indication de fréquence : en effet, de la fréquence on peut toujours déduire la présence ou l'absence, mais l'inverse n'est pas vrai.

Les traitements qui utilisent les indications de présence / absence (modélisées par des booléens, habituellement représentés par des 0 et des 1) ne sont pas pour autant nécessairement inférieurs à ceux qui utilisent les fréquences (représentées par des entiers).

Une raison empirique d'utiliser des indications de présence / absence est lorsque les fréquences effectives ne reflètent pas des variations significatives. Par exemple, si l'on traite des textes très courts, de l'ordre du paragraphe, les reprises lexicales sont quasiment inexistantes, et les quelques rares reprises n'ont pas *sémantiquement* l'impact fort qu'elles obtiendraient *numériquement*.

Une raison théorique de préférer l'information de présence / absence, est lorsque l'objet de la modélisation est lié à l'existence ou la présence d'une expression, indépendamment de son « volume » de réalisation. Ainsi, (Assadi, Bourigault, Gros 1995) étudient les associations entre adjectifs et noms dans les termes techniques, pour en déduire des classes d'adjectifs (paradigmes). Ce qui est significatif pour eux c'est qu'une association Adjectif-Nom soit attestée (existe), ou au contraire ne se trouve pas dans le corpus ; en revanche, il n'y a pas de raison, dans cette étude des possibilités d'associations, de faire de différence entre des termes plus ou moins usités.

Cependant, même dans le cas où l'information cherchée est de l'ordre présence / absence, le passage par les fréquences peut être souhaitable. D'une part, la répétition d'une unité, voire sa présence massive, confirme son attestation : une unité qui n'apparaît qu'une fois ou de façon manifestement rare peut se trouver dans le texte pour des raisons exceptionnelles (une citation, un exemple), qui ne doivent pas biaiser l'analyse. D'autre part, les variations (ou la stabilité) de la fréquence peuvent être utilisées comme un indicateur significatif.

> En ce qui concerne la codification des données, des raisons théoriques nous auraient incités à adopter un critère dichotomique présence - absence (1-0) pour toute une série d'unités. [...] Pour prendre un exemple, dans le profil ou portrait-robot que nous dressons de chaque texte [en vue de caractériser son type], la seule présence d'une 2ème personne du singulier ou d'un déictique temporel devrait suffire à attester l'existence d'un trait caractéristique [...].
>
> Dans l'étape exploratoire, la pratique de l'analyse nous montré qu'un tel critère dichotomique simple ne fonctionne pas à satisfaction dans les conditions de notre étude. Etant donné la difficulté à identifier certaines frontières intratextuelles, l'occurrence unique de telle unité peut être due à la présence dans le passage en question d'un autre type de texte, impossible à détecter si l'on s'en tient à nos critères. En revanche, sa répétition fréquente tend à écarter l'hypothèse d'une erreur due à la délimitation du texte. C'est l'une des raisons qui nous a décidés à relever finalement la *fréquence* des unités observées. Celle-ci a en outre l'avantage de nous indiquer la probabilité d'occurrences (moyenne et dispersion) de chacune de ces unités et de nous permettre d'évaluer le rôle qu'elles peuvent jouer dans l'identification du type de texte.
>
> (Bronckart & *al.* 1985, §V.A.4.1, pp. 70-71)

Il faut ajouter, dans le cas de (Bronckart & *al.* 1985), que les unités observées sont pour la plupart des informations non lexicales (emploi d'un temps, de telle catégorie morphosyntaxique), et donc effectivement susceptibles de fortes fréquences, et de forts écarts de fréquences d'un type de texte à l'autre. Dans une analyse où les unités sont d'ordre lexicale et thématique (comme les travaux de Assadi évoqués juste avant), les éventuelles répétitions peuvent être trop faibles pour être réellement significatives.

L'ensemble des natures de codage numérique

Les nombres d'occurrences des mots dans les textes sont des entiers. Cette évidence a pour conséquence le caractère discontinu, plus exactement discret, des mesures de fréquence. Ce caractère, nous l'avons vu, est particulièrement sensible autour de la fréquence 1. D'autre part, cela souligne le caractère artificiel des réduction homothétiques, qui subrepticement transforment les fréquences (entières) en grandeurs à valeur dans l'ensemble des rationnels (qui a la puissance du continu, au sens mathématique). De fait, si l'on devait « raccourcir » un texte, cela force à des choix : des mots disparaissent, certains restent répétés mais pas d'autres.

Les modèles statistiques sont inégaux devant la prise en compte de la nature des fréquences. La plupart des lois classiques (Poisson, normale ou Gauss) décrivent des variables continues. En se basant sur le modèle hypergéométrique, (Lafon 1980) souligne qu'il adopte un modèle exactement adapté à la distribution des fréquences.

Alors que les manifestations des données semblent du côté du discret et du discontinu, les traitements se laissent souvent mieux penser en termes continus. Ainsi, généralement, les opérations de seuillage (élimination brutale de tout ce qui est en deçà d'une valeur) sont peu satisfaisantes intellectuellement. Le traitement serait donc le passage d'un monde discrétisé (les fréquences par exemples) à une représentation continue.

L'opposition continu / discret ne suffit pas pour décrire les différentes interprétations possibles des mesures au niveau du traitement. On peut en effet distinguer les types suivants :

- *binaire (booléen)* : par exemple, un indicateur de présence / absence d'une unité dans un texte, d'attestation / non attestation d'une unité dans un corpus (Assadi, Bourigault, Gros 1995).
- *symbolique* : par exemple, une lecture des pondérations des termes de métier dans (Bommier, Lemesle 1995), est celle d'un code indiquant quelle(s) année(s) le terme est effectivement paru caractéristique (15 signifie les trois dernières années, 12, les deux dernières, etc.)
- *ordinal* : des expérimentations pour les systèmes de recherche d'information ont été menées, qui utilisent non pas directement les fréquences des mots-clés dans les textes, mais le rang, donnant la hiérarchisation locale à un texte d'importance des mots-clés (Aalbersberg 1994).
- *numéraux discrets* : généralement des entiers naturels, donnant les fréquences de réalisation d'une unité (nombre d'occurrences dans un texte, dans le corpus).
- *numéraux continus* : le passage aux fréquences relatives par exemple fait que l'on travaille avec des opérations mathématiques définies sur l'ensemble des réels.

Penser « 2 » comme un numéral, c'est autoriser à le considérer comme le double de 1, alors que 2 peut être au même plan que 1 si c'est un code symbolique.

De même, un chiffre comme 1 peut aussi bien traduire : l'existence attestée d'une unité, le numéro identifiant une classe à laquelle elle est affectée, sa supériorité dans une certaine hiérarchie, sa position au fil du texte, le fait qu'elle n'apparaisse qu'une fois (hapax), le fait qu'elle n'est pas partagée entre plusieurs textes (elle réalise toutes ses occurrences dans un seul texte), etc. Il est évidemment nécessaire d'expliciter la nature de chaque grandeur, et la signification qui y est associée, pour établir et comprendre les diverses équations.

b) Taille

Ignorer, neutraliser, créditer d'une signification

La question se pose par exemple concrètement en ces termes : j'étudie la littérature française, j'ai bien un corpus disponible, mais il n'est pas aussi fourni pour chacun des quatre siècles que je considère. Par exemple, j'ai un grand nombre d'œuvres du XIX^{ème} siècle, mais nettement moins du XVI^{ème} siècle (moins connu, moins étudié dans les cursus scolaires) ou du Moyen-Age (textes plus difficiles à porter sur support électronique, disparition d'un grand nombre de manuscrits), et le XX^{ème} siècle est intrinsèquement incomplet et hétérogène pour des raisons de tous ordres (les dernières œuvres du siècle n'existent pas encore ; problèmes de droits d'auteur et d'éditeur qui suspendent encore, pendant un délai légal de plusieurs dizaines d'années, la cession publique des textes contemporains ; difficulté à arbitrer les choix de conservation ou de rejet, de séparer œuvres majeures et mineures : le temps n'a pas fait son tri). En connaissance de cet état de fait, faut-il égaliser les différents siècles, ou au contraire considérer leur variation de représentation dans la base comme le reflet d'une réalité effective (l'image d'une certaine conception de la littérature, par exemple) ?

Autre exemple : mon objet d'étude est l'ensemble des thèmes de recherche des chercheurs de la DER d'EDF, et j'ai rassemblé comme corpus de travail l'ensemble des textes descriptifs d'activité (les programmes de recherche rédigés par les chercheurs à l'intention de la Direction). Mon corpus serait naturellement structuré en différents domaines de recherche, délimités par exemple par les Départements (ou les Services). Or ces Départements sont d'effectifs inégaux, et de surcroît les habitudes rédactionnelles, et même la notion de projet, peuvent être très variables : tel Département multiplie les micro-projets, dans tel autre l'accent est mis sur la rédaction des programmes d'activité, qui doit être très fouillée, etc. Faut-il rééquilibrer d'autorité, et mettre sur un pied d'égalité les différents secteurs de l'entreprise (la conception des centrales, l'entretien des lignes, la surveillance des barrages, l'étude des consommations des usagers, etc.) ? Ou faut-il prendre acte des concentrations et des effets de dominance, qui traduisent certaines orientations stratégiques (importance des moyens, intensité de l'activité) ?

Toute la nuance est celle qui sépare l'égalité de l'uniformité. Refuser la partialité ne passe pas nécessairement par l'alignement formel des caractéristiques. La variation de volume elle-même peut être une propriété significative, qu'il ne faudrait pas perdre en s'empressant de la standardiser.

La question concerne la constitution de corpus et leur articulation en sous-corpus, comme dans les deux exemples précédents. Elle concerne également la prise en compte de la longueur des textes, ce que l'on observe à travers l'analyse de plusieurs formules de pondération.

Ignorer

- *Variance*

$$P_u = \frac{1}{T}\sum_{t=1}^{T}\left(F_{ut} - \frac{F_u}{T_u}\right)^2$$

> la variance est [...] un indicateur adapté à des documents de tailles suffisamment importantes et équivalentes. En effet, pour mesurer la dispersion de l'effectif d'un candidat terme [ici : d'une unité] dans un document [ici : dans un texte], il faut que cet effectif soit conséquent. Des documents trop petits (résumés) ne génèrent que des effectifs faibles donc peu dispersés par définition. Dans ce cas la variance ne peut rien discriminer. Les documents doivent être homogènes en taille car sinon les

> expressions contenues dans des textes longs seront privilégiées car elles y ont des effectifs probablement plus importants. (Sta 1997, §6.4.1.2, p. 112)

Ce qui valorise l'unité, c'est d'apparaître avec des effectifs très contrastés, et cela a d'autant plus d'impact si c'est dans peu de documents.

- *Ratio Signal-Bruit (Signal-Noise Ratio)*

L'entropie, ici appelée « bruit » (*noise*), se calcule pour une unité *u* à partir de sa distribution :

$$H(u) = -\sum_{t=1}^{T_u} \frac{F_{ut}}{F_u} \log_2 \left(\frac{F_{ut}}{F_u} \right)$$

Cette formule est une application de la formule de Shannon, en Théorie de l'Information.

On montre que :

$$0 \leq H(u) \leq \log_2(T_u) \leq \log_2(T)$$

H(u) est minimal lorsque l'unité n'apparaît que dans un seul texte (quel que soit alors son nombre d'occurrences) ; *H(u)* est maximal lorsque l'unité est la plus répartie, c'est-à-dire à la fois *(i)* le plus uniformément distribuée, et *(ii)* avec le moins d'occurrences groupées dans des documents. Le maximum extrême $\log_2(T)$ est atteint pour une unité apparaissant dans *tous* les textes, *et* dans chaque texte *une* seule fois.

On définit le « signal » associé à une unité la différence suivante :

$$signal(u) = \log_2(F_u) - H(u)$$

Le signal est toujours positif, nul si et seulement si l'unité est un hapax de texte (*i.e.* quand elle figure dans un texte, elle n'y figure qu'une seule fois). Le signal, qui fournit une pondération intrinsèque, peut être combiné à la fréquence locale dans un texte pour obtenir une pondération contextuelle :

$$P_{ut} = F_{ut} \times signal(u) = F_{ut} \times \left(Log_2\left(F_u\right) - H(u) \right)$$

> A relation clearly exists between noise and term specificity, because broad, nonspecific terms tend to have more even distributions across the documents of a collection, and hence high noise. [...]
>
> In principle, it is possible to rank the index words extracted from the documents of a collection in decreasing order of the signal value. Such an ordering favors terms that distinguish one or two specific documents (the ones in which the high-signal term exclusively occurs) from the remainder of the collection. [...]
>
> The importance, or weight, [P_{ut} of term *u* in document *t* is] analogous to the [inverse document frequency] term weighting function [...].
>
> [...] the signal value does not give optimal performance in a retrieval environment.
>
> (Salton, McGill 1983, §3.3.C, pp. 65-66)

> The [signal value] [...] emphasizes term concentration in only a few documents of a collection and should be used only in order to emphasize precision at the expense of recall.
>
> (Salton, McGill 1983, §3.4, p. 73)

Le Ratio Signal-Bruit a un comportement proche de l'*idf* (*inverse document frequency*), en ce sens qu'il est surtout sensible à la spécificité d'une unité (le fait que l'unité ne soit utilisée que dans peu de textes). En particulier, l'indicateur est favorable à tout hapax de texte, même si c'est un hapax de mot (une unité apparaissant dans un document est valorisée, même si elle n'est attestée qu'une seule fois).

Le Ratio Signal-Bruit a néanmoins des différences potentiellement significatives si l'on a affaire à des textes qui ne sont pas tous courts (autrement dit, que la variation effective des fréquences fait que l'on s'écarte d'un modèle booléen, auquel correspond l'*idf* (Wong, Yao 1992)). Une unité peu courante (qui est plutôt utilisée sous forme de mention, et peu répétée) sera dévalorisée, tout particulièrement si la base comporte des documents longs (ce qui donne l'occasion à d'autres unités d'être répétées). Inversement, une unité reprise souvent dans un document sera valorisée, si bien que, si les documents sont de longueurs irrégulières, et que les uns ne donnent lieu à aucune reprise (ex. : résumé court d'un paragraphe) alors que les autres reprennent abondamment les termes relatifs à leur

thème central, alors l'indicateur favorisera les notions caractéristiques des documents développés aux dépends de celles concernées dans les documents brefs.

Neutraliser

- *Pouvoir informationnel*

C'est encore l'entropie d'une unité qui est estimée, mais d'une façon plus élaborée que le Ratio Signal-Bruit, en prenant en compte la taille des documents via des probabilités conditionnelles (on note ci-après *p(x/y)* la probabilité conditionnelle de *x* sachant *y*). La formule a été proposée par (Fluhr 1977, §III.4-3.2) et ensuite largement reprise et étudiée (Chartron 1988, §VI) (Sta 1997, §6.4.1.5).

Ce que l'on cherche à évaluer, c'est l'information apportée par une unité *u* vis-à-vis du corpus :

$$H\left(Corpus/unité_u\right) = -\sum_{texte_t \in Corpus} p\left(texte_t/unité_u\right) \times \log_2\left(p\left(texte_t/unité_u\right)\right)$$

Cette quantité varie entre 0 (l'unité est dans un seul texte) et $\log_2(T)$ (l'unité est uniformément répartie sur tous les textes) : le pouvoir informationnel d'une unité est d'autant plus grand que $H(Corpus / unité_u)$ est petit (proche de zéro).

Cette formule n'est pas calculable telle quelle : on n'a pas de moyen d'estimer $p(texte_t / unité_u)$. L'astuce trouvée par Fluhr consiste à appliquer la formule de Bayes pour remplacer $p(texte_t / unité_u)$ par une expression en fonction de $p(unité_u / texte_t)$. Chartron poursuit en évaluant $p(unité_u / texte_t)$ par la proportion des occurrences de l'unité *u* par rapport au nombre total d'occurrences dans le texte *t*. On a donc finalement :

$$\hat{H}\left(Corpus/unité_u\right) = \left[\log_2\left(\sum_{t=1}^{T}\frac{L_{ut}}{L_t}\right)\right] - \left[\frac{\sum_{t=1}^{T}\frac{L_{ut}}{L_t}\times\log_2\left(\frac{L_{ut}}{L_t}\right)}{\sum_{t=1}^{T}\frac{L_{ut}}{L_t}}\right]$$

Le nombre d'occurrences de l'unité *u* est constamment relativisé par rapport à la taille du texte considéré : il y a donc bien un effet de neutralisation de la taille sur les variations de fréquence.

La manière que nous avons choisie pour noter la formule détache nettement deux termes. Le premier est le logarithme d'une somme, le second est une moyenne pondérée. On remarque alors certaines valeurs remarquables :

- si l'unité *u* est un hapax de texte (L_{ut} est non nulle pour un seul *t*), alors les sommes se réduisent à un seul terme, et *H* est minimale (nulle). Ceci est vérifié quel que soit par ailleurs le nombre d'occurrences de l'unité *u* dans le seul texte où elle figure, et donc en particulier pour tous les hapax de mot. Cela est perçu comme le principal point faible de la formule : (Fluhr 1977, §III.4-3.3, p. 173) recommande donc d'introduire aussi des critères sémantiques (toute unité qui n'apparaît que dans un seul document n'est pas nécessairement intéressante), et convient que la formule est inadaptée à un corpus comprenant de nombreuses fautes de frappe (« les mots mal orthographiés donnant des graphèmes n'appartenant pas à la langue auraient pour la plupart une entropie nulle et seraient par là-même considérés comme des néologismes très informationnels »).
- si l'unité *u* est équirépartie pour les textes dans lesquels elle apparaît (L_{ut}/L_t vaut soit zéro, soit une constante fixée), alors *H* vaut $\log_2(T_u)$, encore une fois quelle que soit la fréquence relative de *u*. Donc plus *u* est répartie uniformément sur un grand nombre de textes, plus *H* est grand ; *H* atteint sa valeur maximale lorsque *u* est régulièrement distribuée sur tout le corpus. En revanche, comme on peut le souhaiter, *H* est faible si *u* est concentrée dans peu de documents, même si elle est uniformément répartie entre eux. Ce comportement est satisfaisant si l'on considère que la variation proportionnelle à la taille est une forme d'équivalence, et que la comparaison passe par les fréquences relatives (on n'accorde pas de valeur aux fréquences absolues, notamment à la singularité de la fréquence 1).

- Si l'unité *u* est inégalement répartie : *u* apparaît au moins dans un texte. Soit t_0 un texte pour lequel la fréquence relative de *u* est maximale. Alors :

$$\hat{H}\left(Corpus/unité_u\right) \geq \left[\log_2\left(\frac{L_{ut_0}}{L_{t_0}} + \Delta\right)\right] - \left[\log_2\left(\frac{L_{ut_0}}{L_{t_0}}\right)\right]$$

avec

$$\Delta = \sum_{t \neq t_0} \frac{L_{ut}}{L_t}$$

Ceci fait bien apparaître que les valeurs élevées de *H* sont favorisées par les facteurs suivants :
u apparaît dans de nombreux documents,
u a des fréquences relatives fortes,
en particulier *u* apparaît dans des documents courts.
Les unités centrales d'un domaine, bien représentées dans un nombre moyen de documents, s'en trouvent dévalorisées (Chartron 1988, §VI.3, pp. 91-93).

Mis à part le cas des hapax de texte, les limites de cette formule tiennent au fait qu'elle s'intéresse uniquement aux fréquences relatives. En effet, les fréquences absolues peuvent aussi avoir une signification (les hapax notamment). D'autre part, à l'allongement des textes ne correspond pas une augmentation proportionnelles des fréquences : la loi de Zipf montrerait plutôt qu'une bonne part de l'allongement consiste en l'arrivée de nouveaux hapax. La fréquence relative valorise l'impact des textes courts, et donc déséquilibre aussi la représentation (mais dans l'autre sens que l'utilisation des seules fréquences absolues).

- *fonction de pertinence de Lexinet*

(Chartron 1988, §VI.2.1, p. 83 sq.) reprend au groupe SYDO (Laboratoire d'informatique documentaire, Lyon 1), la fonction de pondération d'inspiration statistique suivante :

$$P_u = \sum_{t=1}^{T} \frac{\left(F_{ut} - E_{ut}\right)^2}{{\sigma_{ut}}^2}$$

Espérance et carré de l'écart type sont calculés comme suit :

$$E_{ut} = F_u \times \frac{L_t}{L}$$

$${\sigma_{ut}}^2 = \frac{\left(F - F_u\right) \times F_u \times \left(L - L_t\right) \times L_t}{L^2 \times \left(F - 1\right)}$$

L'espérance E_{ut} est la valeur théorique que prendrait F_{ut} (le nombre d'occurrences de l'unité *u* dans le texte *t*) si l'unité était uniformément répartie sur les textes, proportionnellement à leur longueur. La formule prend en compte la taille de chaque texte (L_t), mais lui fait jouer un rôle très fort : plus un texte est long, plus il est censé avoir des occurrences d'une unité. Inversement, la présence d'une unité dans un texte court est (sur)valorisée.

Plaçons-nous dans le contexte d'un corpus regroupant de nombreux textes (tel que, pour tout texte, sa longueur est négligeable devant le volume de l'ensemble du corpus), et dans le cas d'une unité de fréquence moyenne ou faible (elle n'est qu'une proportion minime du total de toutes les occurrences, et n'apparaît que dans certains textes). Dans ce cas, l'espérance est une bonne approximation du carré de l'écart type, si bien que l'on peut réécrire la pondération ainsi (on numérote les textes en commençant par ceux qui contiennent l'unité concernée) :

$$P_u = \sum_{t=1}^{T_u} \frac{{F_{ut}}^2}{E_{ut}} + \sum_{t=T_u+1}^{T} E_{ur}$$

La seconde somme est négligeable devant la première, puisqu'elle est majorée par F_u (la somme des occurrences prévues sur une partie des textes est une partie du nombre total des occurrences). La première somme met en valeur les unités qui sont répétées dans les textes dans

lesquels elles figurent (le carré amplifie les fréquences supérieures à 1), mais surtout favorise les unités des textes courts. En effet, l'espérance, proportionnelle à la taille du texte, est faible ; et la pondération est inversement proportionnelle à la taille du texte. Autrement dit, pour deux unités qui figurent dans un même nombre de textes, et avec le même nombre d'occurrences dans chacun, si les textes dans lesquels apparaît la première sont deux fois moins longs que ceux de la seconde unité, alors la première unité reçoit une pondération double.

C'est effectivement ce type de comportement qu'observe Ghislaine Chartron sur un corpus de 2 379 documents dans le domaine de l'intelligence artificielle. La fonction de pertinence est moins sensible que la variance au fait que les documents puissent être longs, pour qu'une unité puisse être répétée. En revanche, elle favorise nettement le vocabulaire des textes plus courts que les autres.

Là encore, la prise en considération des longueurs des textes inverse la situation, sans l'équilibrer : certes, les textes longs ne sont plus privilégiés, mais les textes courts gagnent en retour une influence exagérée.

Variation homothétique ou comportement fractal

La solution courante, pour prendre en compte les variations de taille parmi les textes du corpus, est de remplacer la fréquence absolue (nombre d'occurrences de l'unité dans le texte) par les fréquences relatives (nombre d'occurrences de l'unité dans le texte, rapporté à la longueur du texte). C'est en quelque sorte un ajustement d'échelle.

> mesure : [...] le choix des mesures (sur des lignes, entre des repères) est guidé par le principe d'équivalence distributionnelle [...] ; il est un échantillonnage sur un espace probabilisé [..] ; il vise à donner la figure d'ensemble de l'individu (Benzécri & *al.* 1973, § *Indice systématique*)

La similarité dite du « cosinus » met en œuvre l'équivalence distributionnelle, si elle s'appuie sur les fréquences des unités dans les textes. Chaque unité figurant dans le corpus est un axe (une dimension) de l'espace de représentation. Chaque texte est représenté par un vecteur : sur chaque axe, ses coordonnées sont données par le poids dans le texte de l'unité correspondante (on se place dans le cas où le poids est une fonction de la fréquence dans le texte). La similarité entre deux textes est évaluée par le cosinus de l'angle formé par les vecteurs des deux textes. Seule la direction (et donc pas la longueur) des vecteurs intervient pour déterminer la similarité : or la direction est justement déterminée par les unités présentes et leurs fréquences relatives. La similarité est maximale quand les deux vecteurs ont la même direction, c'est-à-dire les mêmes unités avec les mêmes fréquences relatives.

Le principe d'équivalence distributionnelle n'est cependant pas pleinement adéquat pour décrire la répartition des unités dans les textes. En particulier, par la relativité qu'il introduit, il gomme la spécificité de la fréquence 1. Un hapax de mot, qui apparaît dans un texte de longueur l, acquiert exactement la même représentation qu'un hapax de texte, de fréquence 3 dans un texte de longueur triple ($3l$). Or, comme cela a été vu plus haut, le hapax de mot et le hapax de texte ont des valeurs de signification très différentes.

D'autre part, les statistiques elles-mêmes montrent que les fréquences relatives mésestiment les fortes fréquences absolues :

> [Dans le corpus étudié,] la forme *gouvernement* a une fréquence très inégalement distribuée et se trouve être notamment spécifique [...] [des textes] D3 et D5 avec les fréquences locales de 19 et 34. Calculons les fréquences locales relatives : elles sont respectivement de 19/3920 et de 34/7896 soit 0,00485 et 0,00430. Si nous nous fions à cette mesure, nous pourrions dire que gouvernement est légèrement plus employé en D3 qu'en D5. Or les mesures en probabilité [issues du modèle hypergéométrique], qui valent respectivement 21,629 E-03 et 52,009 E-05, nous invitent à inverser ce jugement. De nombreux exemples analogues pourraient être trouvés, ils montrent que si l'on veut comparer l'emploi d'une forme dans deux textes de longueurs inégales, on ne doit accorder qu'une confiance relative à la fréquence relative. (Lafon 1980, p. 152)

D'une manière générale, la diversité des tailles des documents semble souvent ne concerner, dans les formules, que la variable F_{ut} ou L_{ut} (fréquence de l'unité dans le texte, « espace » qu'elle occupe) : car à première vue, un document plus long favorise la répétition des unités. Ignorer les variations de longueur des documents se justifie alors si l'on fait l'hypothèse que les fréquences des unités que l'on veut valoriser varient peu et restent relativement faibles, quelle que soit la longueur du

texte, autrement dit que ces unités ne sont pas de nature à être constamment reprises dans un document, même développé.

> The Signal-Noise Ratio favors terms that are perfectly concentrated in particular documents, and that do not therefore occur in the remaining documents of a collection. On the other hand, the least-useful terms are those which occur evenly in all documents of the collection. In practice, the most concentrated terms tend to be low-frequency terms, whereas the terms with even occurrence characteristics generally exhibit high frequency in a collection. The Signal-Noise Ratio thus exhibits properties somewhat similar to those of the inverse document-frequency factor, and available data show that little is gained by replacing the frequency-based measures of term value with related concepts of information theory. (Salton 1989, §9.3.1, p. 281)

Cette idée, peut-être curieuse à première vue, n'est somme toute pas aberrante. En effet, la loi de Zipf montre bien que les unités répétées sont globalement peu nombreuses (voire minoritaires), et qu'une variation en longueur devrait davantage se traduire par l'apparition de nouvelles unités que par la répétition proportionnelle des unités initiales.

L'observation d'un tel type de comportement pour les unités va directement à l'encontre d'une conception homothétique des fréquences des unités dans les textes. La voie semble plutôt de discerner différents types d'unités en fonction de leur sensibilité à la longueur du texte (en première approximation, la fréquence des items grammaticaux est plus sensible que celle des items lexicaux), et d'affiner la caractérisation de la taille : celle-ci n'est pas nécessairement l'étendue linéaire (en nombre de mots ou de pages), mais peut être pour certains usages le nombre de verbes, la fréquence maximale atteinte par une unité dans le texte (ce que nous avons appelé le calibrage par la fréquence maximale, dans le paragraphe consacré aux fréquences), etc.

> Pour tenir compte de certaines différences dans la longueur des extraits sélectionnés et pour permettre ultérieurement des comparaisons avec des textes de taille différente (nettement plus courts ou plus longs), nous avons calculé des *indices* en rapportant les fréquences observées à un référentiel commun : le total des verbes ou des mots [...]. La référence est le nombre total de verbes pour toutes les unités dont l'occurrence dépend directement de la présence de l'item verbal (les temps, les auxiliaires par exemple) ou plus généralement quand la fréquence de l'unité est fonction du nombre de propositions (les phrases non déclaratives ou l'emphase). [...] [Ainsi,] ces indices ne sont pas biaisés systématiquement par la densité verbale [différente d'un type de texte à l'autre]. (Bronckart & *al.* 1985, §V.A.4.1, p. 71)

Il existe encore une autre manière de faire, qui consiste à n'étudier de chaque texte qu'un extrait de longueur fixée. Ceci repose sur une conception fractale du texte : « je regarde le titre, le sommaire, puis *une* page de contenu (*elles se ressemblent*) ». Autrement dit, un extrait est un échantillon dans lequel se reflète le tout. Plusieurs raisons apparaissent cependant pour ne pas procéder ainsi. Ce parcours partiel du texte correspond à un type de lecture, qui n'est pas nécessairement conciliable avec tout type de textes, ni avec tous les objectifs de lecture (recherche d'un élément précis par exemple). Et concrètement, comment déterminer la taille de référence ? S'en tenir à la taille du texte le plus court (pour ne pas avoir à compléter celui-ci artificiellement) obéit à une logique de nivellement par le bas. Enfin, trancher un morceau du texte est l'introduction pénalisante d'une part d'arbitraire (pourquoi uniquement cette portion ?) et une négation du texte comme unité, unité que le texte réalise dans la perception de son intégralité. Au contraire, l'utilisation du support informatique, avec ses formidables capacités d'enregistrement et de brassage de données, semble livrer un accès nouveau et décuplé au texte intégral : c'est un rendez-vous technologique et scientifique à ne pas manquer.

c) Calculs locaux, calculs globaux

Deux angles de vue complémentaires : intérieur et environnement

Il semble une constante, dans la plupart des modèles de description de corpus de textes, de conjuguer deux types d'indicateurs. Les indicateurs du premier type sont des indicateurs « *intra* » : ils évaluent la consistance interne d'une entité, sa constitution en tant qu'entité autonome. Les indicateurs du second type, complémentaires, sont des indicateurs « *inter* » : ils caractérisent une entité en ce qui concerne son comportement global, ses liens avec les autres entités.

L'équilibre apporté par ce double point de vue explique le succès des graphiques, en représentation plane, dont les deux axes reprennent justement l'*inter* et l'*intra*, baptisés *centralité* et *densité*.

Dans les formules de pondération des unités, il y a également un choix de perspective à faire, dans la place respective à donner à ces deux points de vue, et la manière de les conjoindre. Il y a néanmoins une affinité naturelle forte entre pondération intrinsèque (P_u) et indicateurs *inter*, pondération contextuelle (P_{ut}) et indicateurs *intra*.

- *Variance pondérée*

L'indicateur de la Variance a été étudié plus haut (on a souligné sa sensibilité aux variations de taille des textes du corpus) :

$$P_u = \frac{1}{T}\sum_{t=1}^{T}\left(F_{ut} - \frac{F_u}{T_u}\right)^2$$

La variance pondérée est étudiée par (Chartron 1988, §VI.3), et a la forme suivante :

$$P_u = \sum_{t=1}^{T}\frac{F_{ut}}{F_u} \times \left(F_{ut} - \frac{F_u}{T}\right)^2$$

Ce que l'on veut souligner ici, c'est la différence de « champs » de ces deux formules. La première caractérise une fluctuation de la distribution de l'unité sur tous les textes du corpus. La seconde ne considère que ceux dans lesquels l'unité figure.

La variante pondérée met en valeur les unités dont la fréquence connaît de larges variations (elle hérite donc des mêmes faiblesses que la variance vis-à-vis des corpus de textes courts ou de tailles diversifiées). Elle est moins efficace que la variance simple pour sélectionner les unités spécifiques, qui ne figurent que dans un petit nombre de textes. Ce qu'elle dévalue particulièrement, ce sont les unités qui apparaissent dans beaucoup de textes, et une seule fois (ou avec une petite fréquence) dans chacun (Chartron 1988, §VI.3.2).

Vue focalisée ou vue collective

Soit l'on considère une partie du corpus (un texte ou un ensemble de textes), que l'on veut contraster par rapport au reste, à l'ensemble du corpus, pris comme une totalité : c'est la vue focalisée, qui se centre sur une partie, et laisse l'environnement dans un flou généralisant. Elle reflète une dynamique ascendante : le point de départ, et de référence ultérieure, est une vue locale, une entité particulière. Soit l'on considère le corpus dans sa structure interne, et l'interaction en son sein de différentes parties, quitte ensuite à en déduire des informations pour chacune des parties : c'est la vue collective, qui envisage simultanément un ensemble de parties formant système. La dynamique est alors plutôt descendante : le point de départ est le corpus dans son ensemble, et l'on « descend » dans le détail des parties.

Voici une illustration de la différence vue focalisée vs collective, adaptée de (Bommier, Lemesle 1995). Supposons avoir un corpus qui se subdivise en 9 textes, et deux unités qui se répartissent comme suit, quant aux nombres de positions qu'elles occupent :

	t=1	*t=2*	*t=3*	*t=4*	*t=5*	*t=6*	*t=7*	*t=8*	*t=9*
u=1	20	12	5	0	0	0	0	0	0
u=2	20	5	3	2	2	2	1	1	1

Deux formules sont envisagées, pour mesurer combien une unité est caractéristique d'un texte (en fait, plus généralement, d'une localisation) :

- la formule dite probabiliste simple, qui calcule directement une mesure du degré de présence de l'unité pour un texte déterminé :

$$probabiliste\_simple(u,t) = \frac{L_{ut}/L_t}{L_u/L} = \frac{L_{ut} \times L}{L_u \times L_t}$$

- la formule issue de la théorie de l'information, qui présente l'unité comme caractéristique de la (des) boîte(s) qui apporte(nt) une contribution importante à la somme, sous réserves que le total soit suffisant :

$$\Delta I_u = \left[\sum_{t=1}^{T} \frac{L_{ut}}{L_u} \times \log_2\left(\frac{L}{L_t}\right)\right] - \left[\sum_{t=1}^{T} \frac{L_{ut}}{L_u} \times \log_2\left(\frac{L_u}{L_{ut}}\right)\right] = \sum_{t=1}^{T} \frac{L_{ut}}{L_u} \times \log_2\left(\frac{L_{ut} \times L}{L_u \times L_t}\right)$$

Par la formule probabiliste simple, les unités 1 et 2 ont exactement la même caractérisation vis-à-vis du texte 1, puisque, localement, elles ont la même présence ($L_{1,1}$=$L_{2,1}$=20), et que globalement elles occupent également la même place (L_1=L_2=37). En ce qui concerne la formule issue de la théorie de l'information, les deux unités ont la même contribution à la somme, par contre la somme totale est supérieure pour l'unité 1, dont la distribution est moins étale. Donc l'unité 2, à la différence de l'unité 1, peut être jugée insuffisamment caractérisante pour être caractéristique du texte 1.

Non seulement la seconde formule prend en compte la distribution de l'unité considérée sur tous les textes qu'elle serait susceptible de caractériser, mais aussi sur l'ensemble du corpus, y compris les textes dont elle est absente.

d) Opérateurs et fonctions : combinaisons et transformations

Evaluer la sensibilité réelle

Pour mesurer la similarité entre des noms, en fonction des adjectifs ou des noms avec lesquels ils se construisent pour former des termes composés (ce sont les éléments du *contexte terminologique*), (Assadi 1998, §II.3) examine deux indices de similarité (pour des données binaires) : celui de Jaccard et celui d'Anderberg.

Soit les notations courantes, pour calculer la similarité entre deux noms *i* et *j* :

- *a*, nombre d'éléments communs aux contextes terminologiques des deux noms ;
- *b*, nombre d'éléments dans le contexte terminologique de *i* mais pas dans celui de *j* ;
- *c*, nombre d'éléments dans le contexte terminologique de *j* mais pas dans celui de *i* ;
- *d*, nombre d'éléments recensés dans les contextes terminologiques de tous les noms considérés, et absents des contextes terminologiques de *i* et *j*.

La mesure de similarité de Jaccard est alors :

$$Similarité\_Jaccard(i,j) = \frac{a}{a+b+c}$$

Celle d'Anderberg :

$$Similarité\_Anderberg(i,j) = \alpha \times \left[\frac{1}{2} \times \left(\frac{a}{a+b} + \frac{a}{a+c}\right)\right] + (1-\alpha) \times \left[\frac{1}{2} \times \left(\frac{d}{d+c} + \frac{d}{d+b}\right)\right]$$

Cette dernière formule se comprend ainsi : c'est une combinaison linéaire de deux termes. Le premier mesure la ressemblance 'explicite' par les points communs, le second la ressemblance 'implicite' par les absences communes (ce qu'ils n'ont ni l'un ni l'autre). Le paramètre α règle l'équilibre entre ressemblance explicite et ressemblance implicite : s'il vaut 1 (resp. 0), on ne considère que les ressemblances explicites (resp. implicites) ; et pour une valeur intermédiaire entre 0 et 1, plus la valeur est proche de 1, plus la ressemblance explicite est influente et la ressemblance implicite secondaire.

Mis à part le coefficient en α dont nous venons de voir le rôle, chacun des deux termes est une moyenne arithmétique. C'est la moyenne de la proportions des caractères communs pour chacun des deux noms. Le premier terme effectue ce calcul sur les caractères présents (proportions prises dans les contextes terminologiques de chacun des noms), le second sur les caractères absents (proportions hors des contextes terminologiques de chaque nom). Prendre en compte ces caractères absents, c'est leur attribuer un sens : par exemple, les deux excluent tous deux tel élément, qui ne peut faire partie de leur contexte terminologique ; ils ont la même incompatibilité.

La question est ici celle de la sensibilité réelle de chaque partie de la formule :

- si α est fixé très proche de 1 par exemple, seul le premier terme joue un rôle ;
- si le contexte terminologique d'un nom est une petite partie de l'union de tous les contextes terminologiques, alors *d* est grand devant *b* et *c* (et *a*), et le second terme se comporte comme une constante : là encore, il ne joue aucun rôle pour caractériser la similarité.

Dans ces deux cas, la formule complexe d'Anderberg se rapproche de la formule plus simple de Jaccard, et fournit dans les faits des valeurs de similarité suivant la même interprétation : deux noms sont d'autant plus semblables qu'ils ont un fort recouvrement réciproque de leurs contextes terminologiques.

Transposons ceci à une mesure de similarité entre deux textes d'un corpus, en fonction des unités que chacun comporte : si chaque texte n'utilise qu'une très petite partie de toutes les unités présentes dans le corpus, alors la sensibilité réelle du second terme d'Anderberg est nulle, et l'interprétation que l'on doit donner aux valeurs de similarité n'a à considérer que le premier terme.

L'analyse de la sensibilité réelle des formules est un moyen d'alléger les calculs, de simplifier le modèle mathématique, et de clarifier l'interprétation en la focalisant sur les éléments qui jouent un rôle effectif.

Logarithme

Le logarithme se définit comme transformant la multiplication en addition (le logarithme d'un produit est égal à la somme des logarithmes des facteurs). Ainsi, les différences de logarithmes mesurent un facteur multiplicatif (soit une comparaison relative, invariante par homothétie), plutôt qu'une différence absolue.

Le logarithme, appliqué à des grandeurs positives supérieures à 1, est une transformation monotone, qui atténue progressivement et vigoureusement les valeurs à mesure qu'elles croissent. Ainsi, dans la combinatoire de formules de pondérations explorées en *Information retrieval* (Hersh & *al.* 1994), le logarithme est typiquement applicable à l'*inverse document frequency*, ou à la fréquence de l'unité dans le texte. L'*idf* ainsi corrigée, au lieu d'avoir une plage de variation entre 1 et 2 000 (si 2 000 est le nombre de documents dans la base), varie entre 1 et 12, ce qui est beaucoup plus raisonnable : une unité qui figure dans deux ou trois documents est dix fois plus discriminante qu'une unité apparaissant dans une majorité des documents (ce qui est plus équilibré que mille fois plus). En outre, la sensibilité est avivée pour les écarts sur un petit nombre de documents, alors qu'elle est très atténuée sur les grands ensembles, ce qui correspond bien à l'intuition : la différence qualitative perçue entre une unité qui apparaît dans 2 documents et une qui apparaît dans 4, est beaucoup plus forte que la différence perçue entre une unité qui apparaît dans 50 documents et une qui apparaît dans 52. En ce qui concerne l'application du logarithme à la fréquence de l'unité dans le texte, les effets recherchés sont sensiblement les mêmes : atténuation des fortes fréquences, et sensibilité accrue sur les faibles fréquences (on remplace F_{ut} par $(1 + \log(F_{ut}))$, pour ne pas éliminer les unités de fréquence 1). Concrètement, on veut marquer la différence significative qu'il y a entre une unité apparaissant une fois (hapax) et une unité apparaissant deux ou trois fois (elle est « confirmée » par sa répétition). En revanche, qu'une unité apparaisse 10 ou 20 fois, ce que l'on veut retenir, c'est qu'elle apparaît beaucoup, peu importe exactement combien de fois, et l'écart dans ces gammes de fréquences n'est plus autant significatif.

Il y a aussi des usages « rigoureux » du logarithme, qui est alors introduit non pas pour son intérêt heuristique, mais suite à une démonstration mathématique. Le logarithme (base 2, mais tout autre logarithme est équivalent à une constante multiplicative près) est central en théorie de l'information, notamment pour la formule de Shannon. Il vient du nombre d'étapes nécessaires pour localiser, par dichotomies successives, un élément dans un ensemble (Volle 1985, §III). Par exemple, dans un ensemble de 16 éléments, $\log_2(16) = 4$ divisions suffisent pour cerner un élément unique (la première division fait passer de 16 à 8 éléments, la seconde de 8 à 4, la troisième de 4 à 2, la dernière de 2 à 1). La procédure demande 4 unités d'informations, une unité par étape, à laquelle il faut choisir la bonne moitié.

L'exponentielle, qui est la fonction inverse du logarithme (népérien), est quant à elle requise dans l'expression de la loi de Poisson. La loi de Poisson est un modèle statistique pour décrire les

événements rares : l'apparition, dans des textes, d'un mot peu courant, peut être approximée par cette loi, de préférence à la loi normale (Muller 1973, §20).

Homogénéiser : Rapport

Les fractions sont l'expression typique des taux (proportions). Elles ont une valeur de *normalisation* : si l'on a affaire à de grandeurs positives, et que le dénominateur est un majorant de la variable au numérateur, le rapport est *borné*, et prend une valeur « *sans dimension* » entre 0 et 1, décontextualisée, directement comparable avec toutes les autres mesures du même type, puisque sur la base d'une référence commune. Si ce dénominateur est non seulement un majorant, mais plus spécifiquement le maximum (possible), ou le total (réalisé), le rapport peut être interprété comme une évaluation d'une *probabilité*.

Fusionner : Moyennes

Pour obtenir une valeur synthétique et centrale de n mesures homogènes $x_1, x_2,... x_n$, on peut calculer leur moyenne. Il y a en fait plusieurs types de moyennes, dont il convient de souligner les propriétés respectives :

- la moyenne arithmétique :

$$Moyenne\_Arithmétique(x_1, x_2, \ldots, x_n) = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- la moyenne géométrique (pour des grandeurs positives) :

$$Moyenne\_Géométrique(x_1, x_2, \ldots, x_n) = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n} = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

- la moyenne harmonique :

$$Moyenne\_Harmonique(x_1, x_2, \ldots, x_n) = \frac{1}{Moyenne\_Arithmétique\left(\frac{1}{x_1}, \frac{1}{x_2}, \ldots, \frac{1}{x_n}\right)}$$

La moyenne arithmétique est invariante pour toute modification des valeurs individuelles $x_1, x_2,... x_n$ qui laisse inchangée la somme totale. En particulier, le système $x_1+k, x_2-k, x3,... x_n$ a la même moyenne arithmétique que le système initial. La moyenne géométrique est à l'inverse très sensible au fait que $x_1, x_2,... x_n$ aient des valeurs proches les unes des autres ou non. La moyenne géométrique est d'autant plus faible que, à somme constante, les $x_1, x_2,... x_n$ ont des valeurs disparates. On montre par exemple très simplement la décroissance de la moyenne géométrique de deux valeurs lorsqu'elles s'écartent :

$$x^2 > (x-k)\,(x+k) = x^2 - k^2$$

Des considérations d'implémentation peuvent faire préférer la moyenne arithmétique. La moyenne géométrique « passe » par une valeur intermédiaire, le produit des x_i, qui dans certains cas risque de sortir des limites du calculateur. Si l'on fait une moyenne géométrique de nombreuses valeurs strictement comprises entre 0 et 1, cela exige un calcul d'une très grande précision pour ne pas avoir un résultat faux (égal à zéro en l'occurrence). S'il s'agit d'une moyenne géométrique de valeurs supérieures à 1, la multiplication peut faire atteindre des ordres de grandeurs non traités par la machine. Enfin, le calcul de la racine carré est une opération plus complexe que les opérations requises pour la moyenne arithmétique.

Les moyennes fournissent une valeur comprise entre la plus petite valeur des x_i et la plus grande. Mais aucune des formules n'assure de trouver une valeur parmi celle des x_i de départ. De même, aucune n'assure que la synthèse est un entier, si les x_i sont des entiers naturels quelconques. Ceci est l'observation d'un état de fait, qui peut tout à fait être rectifié si nécessaire par une fonction d'arrondi (partie entière, ou arrondi au x_i le plus proche, etc.).

La *médiane* se présente comme un moyen très simple d'avoir une valeur « centrale » pour les $x_1, x_2,... x_n$, et qui soit un des x_i : si l'on ordonne les x_i du rang 1 au rang n selon leur valeur, alors la médiane est le x_i du rang du milieu ((n+1)/2 si n est pair, le rang juste au dessus ou juste au-dessous,

selon la convention adoptée, si *n* est impair). Autre indicateur synthétique qui renvoie une des valeurs parmi les x_i : le *mode*, qui correspond à la valeur la plus prise par les x_i. Le mode n'a de sens que pour des x_1, x_2,... x_n dont les valeurs reflètent certains types de distribution, et sont suffisamment « denses » pour que des valeurs soient répétées, et ce de façon significative.

Particulariser : facteurs multiplicatifs et puissances

Les moyennes accordent une égale importance à chacune des grandeurs x_1, x_2,... x_n. Les moyens de relativiser le rôle joué par chacune des variables sont :

- des coefficients multiplicatifs, pour la moyenne arithmétique[10] :

$$Arithmétique\_Pondérée(x_1, x_2, \ldots, x_n) = \frac{\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n}{\alpha_1 + \alpha_2 + \ldots + \alpha_n} = \frac{\sum_{i=1}^{n} \alpha_i x_i}{\sum_{i=1}^{n} \alpha_i}$$

- des puissances, pour la moyenne géométrique :

$$Géométrique\_Pondérée(x_1, x_2, \ldots, x_n) = \sqrt[(\alpha_1+\alpha_2+\ldots+\alpha_n)]{x_1^{\alpha_1} \times x_2^{\alpha_2} \times \ldots \times x_n^{\alpha_n}} = \left(\prod_{i=1}^{n} x_i^{\alpha_i} \right)^{1/\sum_{i=1}^{n} \alpha_i}$$

Il est possible aussi de proposer des combinaisons s'inspirant des distances de Minkowski :

$$Minkowski(x_1, x_2, \ldots, x_n) = \sqrt[q]{x_1^q + x_2^q + \ldots + x_n^q} = \left(\sum_{i=1}^{n} x_i^q \right)^{1/q}$$

q est un entier ; plus q est grand, plus la valeur résultante est grande et proche des plus grands x_i ; lorsque q tend vers l'infini, l'opération est équivalente à la sélection du x_i de valeur maximale.

Contrôler la propagation

Les mesures de fusion peuvent ajuster l'importance accordée à chaque élément, mais obligent néanmoins à considérer tous les éléments. Si quelques éléments apportent une contribution mauvaise, alors l'ensemble est dévalorisé.

Dans les modèles étudiés ici, ces mesures de fusion sont typiquement utilisées pour construire la représentation d'un texte, à partir de celles de ses unités. Les importances relatives des unités sont exprimées par des pondérations. Or ce passage, des unités isolées au texte dans son entier, n'est pas une modélisation adéquate de la contextualisation. En effet, toute unité n'est pas *indifféremment* en interrelation avec *toute* autre. C'est oublier, au plan syntagmatique, l'effet des zones de localité (deux unités composant un syntagme ont une relation beaucoup plus manifeste que deux unités séparées par un flot de paragraphes). Au plan paradigmatique, il est clair aussi qu'il existe des affinités sémantiques plus ou moins stabilisées et plus ou moins constructives de sens. En définitive, alors que les mots-clés sont des unités *dé*contextualisées, la combinaison générale des mots-clés est *sur*contextualisante.

---

[10] (Dupuy 1993) fait ainsi volontiers usage de combinaisons linéaires, opérant hélas un amalgame informe et arbitraire d'indicateurs initialement prometteurs. Par exemple, il additionne, modulo quelques facteurs multiplicatifs :
- la taille du syntagme dont les occurrences sont étudiées,
- sa fréquence relative dans le texte considéré,
- sa valeur (codée de 1 à 6, selon sa proportion de mots outils vs de lexique thématique),
- sa position (qui reçoit un score de 3 à 15, selon que le syntagme apparaît dans le titre, la première ou la dernière phrase du texte, ou en début ou en fin de paragraphes)
- sa portée (écart moyen en nombre de phrases entre les occurrences),

et l'on obtient... le *poids itératif* (Dupuy 1993, pp. 109-113).
La synthèse opérée ici par l'addition nous semble pour le moins brutale et manque totalement de justification, Jean-Philippe Dupuy n'entreprenant pas lui-même de la défendre. La présence de ce type de formules, abusives et gratuites, est notre plus grand regret par rapport à cette thèse de Dupuy, par ailleurs débordante d'idées.

Il y aurait donc lieu de limiter l'impact des pénalisations, pour que les interactions significatives, mais locales, entre quelques mots-clés ne soient pas étouffées. C'est en quelque sorte la tactique de l'insubmersible : il faut cloisonner les parois du navire, de sorte qu'un accroc à un endroit n'ait qu'une incidence limitée, et ne fasse pas systématiquement couler le bâtiment.

Parmi les propositions en fin de ce chapitre, le modèle exposé pour la gestion de l'interaction entre les unités propose un mode de combinaison des unités caractérisantes qui tient compte de cette contextualisation nuancée.

L'union fait la force : complémentarité des indicateurs

Les formules, dans leur diversité, décrivent des propriétés différentes et sont complémentaires (Sta 1997, §6.5.2, pp. 124-125). Aussi il se montre plus judicieux non pas de chercher « la » meilleure formule, mais d'utiliser plusieurs indicateurs et de combiner leurs résultats.

e) Ecarts et exceptions

Les pratiques de classification montrent que, quel que soit le point de vue choisi, la grille descriptive qu'on se donne, il reste toujours une partie de la réalité à décrire qui ne trouve pas sa place de façon adéquate dans le classement. L'analyste doit alors choisir une tactique. Ou bien il fait disparaître ces éléments, en les sortant de la réalité à décrire. Ou bien, il les intègre malgré tout chacun dans une classe, en essayant de limiter au maximum les distorsions résultantes. Ou encore, il les isole dans une classe « divers », qui n'a en fait de classe que le nom, car elle est constituée d'éléments dépareillés et n'a donc pas de cohérence interne.

Les procédures visant à produire une vue synthétique sur des données adoptent également un certain comportement vis-à-vis des éléments originaux, particuliers et rares, dont il faut prendre conscience préalablement à leur emploi. Dans la plupart des cas, la synthèse gomme les points aberrants : une simple moyenne par exemple ne reflète pas un écart isolé ; à l'inverse, un indicateur qui retiendrait le minimum et le maximum de la valeur prise par une variable est très sensible à des valeurs qui sortent de l'ordre de grandeur habituel. Un instrument comme l'analyse factorielle est un peu différent : elle ne fait ressortir ni des particularités isolées, ni un comportement moyen, mais elle s'intéresse à valoriser les singularités significatives, autrement dit des écarts typiques d'une partie non négligeable des données. La valeur de discrimination (présentée au paragraphe ci-après) fait elle aussi ressortir des éléments originaux (des unités qui distinguent un petit nombre de textes), mais pas « trop » originaux (une unité qui n'apparaît que dans un ou deux textes est moins utile à la description du corpus dans son ensemble et des rapports entre les textes). C'est en cela que la valeur de discrimination a pu apparaître comme une alternative intéressante à l'*idf* (*inverse document frequency*), qui elle met au premier plan les singularités ponctuelles (les unités qui ne figurent que dans un seul document reçoivent le score le plus fort).

f) Au fait, que veut-on bien savoir au départ ?

- *Valeur de discrimination (Term-Discrimination Value)*

L'hypothèse sous-jacente est qu'une unité est d'autant plus intéressante qu'elle caractérise un certains nombre de textes par rapport aux autres (dans lesquels elle ne figure pas). Autrement dit, elle est à la fois sélective, et cependant pas trop rare, pour être efficace et rentable.

L'astuce consiste à calculer la densité de l'espace de représentation des textes (ou documents), ce qui revient à savoir déterminer la distance moyenne (le contraste moyen) entre deux textes quelconques. Alors, une unité est d'autant plus intéressante que ce contraste est fortement augmenté quand on la prend en compte. De façon imagée, l'effet de l'unité est d'écarter les textes et d'élargir l'espace de description, ce qui permet d'y voir plus clair, d'avoir une représentation moins grossière, confuse et brouillée.

Typiquement, les unités ainsi sélectionnées sont celles qui apparaissent dans un nombre moyen de textes. En effet, celles qui apparaissent dans beaucoup de textes ont une valeur de discrimination négative (elles sont un facteur de ressemblance générale entre les textes, donc le

contraste est diminué quand on les prend en compte). Inversement, pour les unités qui apparaissent dans un trop petit nombre de textes, leur ajout ou leur retrait n'a pas d'incidence sensible sur le contraste d'ensemble, si bien que leur valeur de discrimination est nulle. C'est là une différence importante par rapport à l'*idf* : l'*inverse document frequency* est d'autant plus grande que l'unité ne figure que dans peu de textes. Pour la valeur de discrimination, lorsque le nombre de textes contenant l'unité augmente, la valeur commence par croître avant de décroître (Salton 1989, §9.3.2, p. 283).

Le zéro constitue une valeur charnière naturelle, qui rend moins arbitraire la fixation d'un seuil (Sta 1997, §6.4.1.3, p. 113).

On peut trouver le détail des formules dans (Salton 1989, §9.3.2, pp. 281-284), et une présentation illustrée dans (Sta 1997, §6.4.1.3, pp. 112-113).

Il y a nettement quelque chose de l'ordre de la détermination du local par le global, dans cette approche : en effet, c'est en raison de son effet dans le contexte global dans la caractérisation des textes du corpus, que l'unité se voit valorisée ou non. Mais on rencontre aussi directement le cercle herméneutique, car comment sont mesurées les distances entre les textes, sinon déjà en utilisant les unités (figurant localement dans les textes) et leurs éventuelles pondérations ?

Il y a donc un aspect intrinsèquement cyclique, dans le calcul de la Valeur de discrimination. Grosso modo,

- pour calculer la pondération d'une unité, il faut connaître les distances entre textes ;
- et pour déterminer les distances entre textes, on ne dispose généralement pas d'autres moyens que de faire un calcul qui utilise les unités composant chaque texte ;
- enfin, les meilleures formules de distance classiques font intervenir une pondération des unités –la boucle est bouclée.

- *Densité locale*

L'indicateur de densité locale est proposé par Sta (Sta 1997, §§6.4.1.4, pp. 113-114) pour identifier les termes en raison de leur spécificité thématique. L'idée est qu'une unité est intéressante à repérer si elle est typiquement associée à une certaine thématique, et cela se traduit en vérifiant que les textes qui la contiennent se ressemblent (ils ont le même vocabulaire).

La densité locale d'une unité se calcule donc en déterminant la moyenne des similarités entre les textes qui comportent l'unité. Ici l'analogie avec le principe de calcul de la Valeur de Discriminance ressort clairement : le poids d'une unité suppose connues les similarités entre textes, alors que celles-ci sont habituellement calculées en fonction des pondérations.

Cette mesure est séduisante pour sa dimension contextuelle, que l'on pourrait étudier davantage : densité locale calculée non plus à partir des ressemblances des textes dans leur entier, mais de celle des paragraphes, des phrases... Chez (Sta 1997), le choix du document tient vraisemblablement à la taille régulière et pas trop grande (une à deux pages) des textes de son corpus, et au fait que le texte représente bien une unité thématique (alors que à l'échelle de la phrase les interactions syntaxiques sont beaucoup plus présentes).

- *Modèles probabilistes*

Dans le domaine de la recherche d'information, des pondérations des mots-clés sont évaluées en fonction de la probabilité d'un mot-clé à sélectionner un sous-ensemble de documents intéressants (Salton 1989, §9.3.3). Cette approche présente de limites manifestes :

- elle se base sur des jugements de pertinence figés, associant des paires requête (mot) - document dans l'absolu.
- non seulement l'utilisation de pertinences *a priori* est contestable dans son principe, mais elle pose aussi de sérieuses difficultés pratiques, car il faut disposer d'un très grand nombre de tels jugements pour avoir des approximations acceptables des probabilités.
- chaque mot clé est considéré indépendamment, hors contexte ; la prise en compte de l'interaction entre les mots-clés, dont le bien-fondé est admis, est cependant exclue (ou fortement limitée) pour des raisons de complexité des calculs. Le modèle ne s'adapte donc pas bien à la réalité linguistique.

En ce qui concerne la diffusion ciblée, un filtrage des termes a été expérimenté, en se basant sur un ensemble d'envois validés ou invalidés (Sta 1994). Il s'agissait donc d'une forme

d'apprentissage à partir d'un échantillon test. Cette expérience n'a pas été poursuivie, car l'obtention de validations sur les envois ne peut être qu'exceptionnelle (charge de travail induite, participation volontaire des utilisateurs *a priori* faible et irrégulière). Elle peut se justifier ponctuellement pour la mise au point d'un prototype, mais non régulièrement pour l'entretien d'une application en exploitation.

C. PROPOSITIONS

1. Moyens

a) Dans le modèle proposé pour DECID : passage de la description à la caractérisation

Dans un chapitre précédent, ont été présentés trois natures d'unités :

- les unités élémentaires, qui sont les *types* correspondants aux *occurrences* telles quelles relevées dans les textes ;
- les unités descriptives, qui sont définies à partir des unités élémentaires, et constituent le « dictionnaire » des unités utilisables pour caractériser les textes ;
- les unités caractérisantes, qui sont les unités descriptives affectées à un texte : dans le contexte de ce texte, elles sont enrichies par des indications sur leur mode de réalisation et d'attribution.

Les pondérations intrinsèques, qui sont des fonctions d'une unité et d'un corpus, sont des grandeurs associées aux unités descriptives. Les pondérations contextuelles, qui sont des fonctions d'une unité et d'un texte dans un corpus, concernent les unités caractérisantes. Les similarités texte - texte sont des opérations entre deux ensembles d'unités caractérisantes.

b) Propriétés textuelles potentiellement caractérisantes

Le contenu de ce paragraphe est un inventaire de travail de grandeurs significatives calculables automatiquement. Chaque grandeur reflète une propriété, une appréciation par nature qualitative. Il y aurait de multiples manières d'en donner une expression chiffrée : aucune n'est optimale et définitive, même si certaines manières sont préférables à d'autres, en ce sens que leur interprétation correspond mieux à la propriété.

Chaque mesure est esquissée succinctement : *(i)* intuitivement, par son nom, évocateur de la propriété visée ; *(ii)* opérationnellement, par une proposition sur la manière de la calculer dans le modèle construit dans le cadre de cette recherche ; *(iii)* une illustration simple est donnée par une formule, utilisant autant que possible les grandeurs de base (recensées plus haut et dotées d'une notation conventionnelle unifiée). Le but de l'exemple illustratif est d'aider à la compréhension de la propriété visée, et en général il ne traduit pas exactement la proposition forgée dans le cadre du modèle. Pour rester simple et général, il ne tient pas non plus toujours compte des limites soulignées précédemment pour certaines formules. Mais on peut ainsi s'en tenir aux notations déjà introduites, montrer que la propriété est à la convergence des différents calculs envisagés, donner une formule simple et facilement compréhensible sans renoncer à une formulation plus élaborée dans le cadre du modèle.

Plusieurs de ces indicateurs ont été mis en œuvre pour le calcul de représentations synthétiques de groupes de textes dans un corpus (Bommier, Lemesle 1995) : un groupe de textes n'est pas perçu comme la juxtaposition ou le cumul des textes qui le composent, mais comme une récapitulation de ses caractéristiques globales et de certains points saillants. Ce calcul est appliqué à la construction automatique des profils des Groupes, Départements et Services d'EDF-DER, à partir de l'ensemble des textes rédigés par les chercheurs. Ainsi, un Département n'est pas la « somme » des textes des personnes du Département, mais un récapitulatif des dominantes qui ressortent globalement de ces textes. Ces profils des structures de la DER sont intégrés à l'application de diffusion ciblée, permettant ainsi de repérer non seulement des personnes, mais également des équipes plus larges de la DER, concernées par un sujet.

Ce travail est loin d'être abouti, il n'est pas encore dans une phase de validation expérimentale, et moins encore dans une phase conclusive. Pour autant l'expression et l'organisation des différents types de mesure représentent un premier acquis important. C'est la base nécessaire pour repenser et affiner les formules de calcul dans l'application de diffusion ciblée (pondérations,

similarités). Ces éléments pourraient également être repris dans d'autres recherches impliquant des calculs sur données textuelles.

Mesures des échelles

- *effectif* : pour un segment qui n'est pas une orientation, nombre de divisions.
  ex. : *T*, le nombre de textes, comme effectif du corpus.
- *longueur* : pour un segment qui est une orientation, nombre de divisions.
  ex. : L_t, le nombre de positions dans le texte *t*, comme longueur du texte *t*.
- *spectre* : pour une unité descriptive, nombre d'unités descriptives de base qui entrent dans sa composition (directement ou indirectement).
  ex. : N_u, le nombre de formes différentes que peut prendre l'unité *u* dans le corpus.
- *étalon* : une (ou plusieurs) grandeur(s) de référence, adaptée à la variable que l'on veut mesurer, typiquement le maximum théorique ou observé de la valeur de la variable.
  ex. : la fréquence maximale d'une unité dans le texte *t* :

$$étalon_{F_{ut}}(t) = \max_{u \in Vocabulaire\_de\_description} \left(F_{ut}\right)$$

Mesures de l'adéquation et de la manifestation

- *présence* : pour une unité caractérisante (donc attribuée à un segment), nombre d'occurrences d'unités descriptives de base contribuant à sa réalisation dans le segment.
  ex. : F_{ut}, fréquence (nombre d'occurrences) de l'unité *u* dans le texte *t*.
- *déploiement* : pour une unité caractérisante, nombre d'unités descriptives de base contribuant à sa réalisation dans le segment.
  ex. : N_{ut}, le nombre de formes différentes que prend l'unité *u* dans le texte *t*.
- *prise en charge* : pour un texte et un univers (un ensemble d'unités descriptives disponibles), taux d'unités élémentaires non reconnues.
  ex. : dans une représentation où il ne peut y avoir plusieurs unités occupant une même position dans le texte, et où les unités, pour être comptabilisées, doivent être reconnues à partir d'un vocabulaire de référence, alors la prise en charge d'un texte t pourrait être évaluée par le taux de positions non occupées par une unité reconnue :

$$Prise\_en\_Charge\ (t) = (L_t - F_t) / L_t$$

Mesures de l'intensité

- *redondance* : nombre d'occurrences de l'unité élémentaire qui contribue à la réalisation de l'unité caractérisante et qui est la plus répétée dans le texte ; une valeur de redondance peut aussi être attribuée à l'unité descriptive, en faisant la moyenne des redondances dans les différents textes où l'unité est présente.
  ex. : F_{ut} / N_{ut}, nombre moyen d'occurrences d'une forme de manifestation de l'unité *u* dans le texte *t*.[11]
- *visibilité* : moyenne de la *présence* de l'unité, dans les textes où elle figure.
  ex. : F_u / T_u, fréquence moyenne de l'unité u dans les textes du corpus où elle est présente.

[11] (Dupuy 1993) propose une mesure analogue, sous le nom de *dispersion* :
« On appellera *dispersion* [ou *variance*] lexicale le rapport entre la fréquence d'une isotopie et le nombre de vocables lexicalement différents qui y participent [...]. L'intérêt de ce paramètre est qu'il évalue la variance lexicale de l'isotopie [...] qui, [procédant] à la fois de l'itération (un même champ lexical) et de la différence (des lexèmes différents entrant en opposition), est une manifestation privilégiée du sens. » (Dupuy 1993, p. 368)

- *éclat* : moyenne de la présence de l'unité, dans les 5 % des textes où elle est le plus présente (c'est-à-dire : dans le texte où elle est le plus présente si elle est présente dans 1 à 20 textes, dans les deux textes où elle est le plus présente si elle est présente dans 21 à 40 textes, etc.)
  ex. : nombre maximal d'occurrences de l'unité *u* dans un texte du corpus :
$$Eclat(u) = \max_{t \in Corpus} \left(F_{ut} \right)$$

- *contraste* : variance de la présence de l'unité descriptive, sur l'ensemble des textes dans lesquels elle est présente.
  ex. : variance du nombre d'occurrences de *u*, dans les textes du corpus où elle figure :
$$Contraste(u) = \frac{1}{T_u} \sum_{t/F_{ut}>0} \left(F_{ut} - \frac{F_u}{T_u} \right)^2$$

- *monotonie* : pour une unité descriptive, une mesure de la corrélation entre la présence en tant qu'unité caractérisante dans un texte et la longueur du texte.
  ex. : variance de la fréquence relative de *u*, dans les textes du corpus où elle figure (on prend l'expression de la variance centrée sur zéro, pour simplifier les calculs) :
$$Monotonie(u) = \left[\frac{1}{T_u} \sum_{t/F_{ut}>0} \left(\frac{F_{ut}}{L_t} \right)^2 \right] - \left[\frac{1}{T_u} \sum_{t/F_{ut}>0} \frac{F_{ut}}{L_t} \right]^2$$

- *magnitude* : pour une unité caractérisante, présence relative aux unités caractérisantes du même texte.
  ex. : rapport entre le poids de l'unité *u* dans le texte *t*, et la moyenne des poids des unités dans *t*, en prenant un poids de type *tf.idf* (term frequency, inverse document frequency) :
$$Magnitude(u,t) = \frac{F_{ut}/T_u}{\sum_{u' \in t} \frac{F_{u't}}{T_{u'}}} \times N_t$$

Mesures de répartition

- *participation* : pour une unité descriptive, nombre de textes dans lesquels elle apparaît.
  ex. : T_u, nombre de textes où figure l'unité *u*.
- *influence* : pour une unité descriptive, nombre de textes où elle est dominante.
  ex. : F_u / T_u, fréquence moyenne de l'unité u dans les textes où elle apparaît.
- *perméabilité* : pour une unité descriptive, proportion, dans les textes dans lesquels elle apparaît, de ceux pour lesquels elle a une présence moyenne ou faible (présence inférieure à sa présence moyenne et à la moitié de sa présence maximale).
  ex. : nombre de textes dans lesquels l'unité u apparaît avec une fréquence strictement inférieure à sa fréquence moyenne, c'est-à-dire de textes t tels que $F_{ut} < (F_u/T_u)$
- *discriminance* : mesure de l'information (au sens de la théorie de l'information) apportée par une unité descriptive *u* par rapport à des découpages internes du corpus (par exemple, répartition des dans boîtes).
  ex. : $\log(T/T_u)$
- *spécificité* : pour une unité caractérisante *u* d'un texte *t*, mesure de la réalisation de *u* dans *t* par rapport à toutes les autres réalisations de *u* dans les autres textes du corpus.
  ex. : forme d'écart réduit :
$$Spécificité(u,t) = \frac{F_{ut} - \frac{F_u}{T}}{\sqrt{F_u}}$$

- *représentativité* : présence relative de l'unité caractérisante *u* dans le texte *t*.
  ex. : rapport entre le poids de l'unité *u* dans le texte *t*, et la somme des poids des unités dans *t*, en prenant un poids de type *tf.idf* (term frequency, inverse document frequency) :

$$représentativité(u,t) = \frac{F_{ut} / T_u}{\sum_{u' \in t} \frac{F_{u't}}{T_{u'}}}$$

- *couverture* : étendue relative de l'unité caractérisante *u* dans le texte *t*.
  ex. : proportion de paragraphes du texte *t* entre les premières et les dernières occurrences de l'unité *u*.

Mesures de localisation

- *ancrage* : localisation(s) où la présence de l'unité descriptive *u* est la plus forte.
  ex. : pour l'unité *u*, texte(s) *t* où F_{ut} est maximale (*u* fixée, *t* décrivant le corpus).
- *rayonnement* : localisation(s) où la présence de *u* est significative.
  ex. : pour l'unité *u*, l'ensemble des textes où elle est présente.

Mesures de rythme - perception

Ces mesures se réalisent sur des segments de type orientation, par exemple sur un texte, ou en faisant la synthèse de *n* valeurs sur *n* textes.

- *portée* : moyenne des longueurs minimales entre unités élémentaires, correspondant à des unités descriptives de base différentes, et concourant effectivement à la réalisation de l'unité descriptive.
  ex. : moyenne, sur l'ensemble des occurrences de l'unité *u* dans les textes où elle est répétée, du nombre d'occurrences d'autres unités entre chaque occurrence de *u* et l'occurrence de *u* qui lui est la plus proche.
- *répartition linéaire* : mesure de la répartition régulière vs par groupes espacés des occurrences concourant à la réalisation de l'unité descriptive.[12]
  ex. : rapport entre l'écart-type et la moyenne de la distance entre deux occurrences successives de l'unité *u* (Noe 1985) (voir aussi les travaux, en lexicométrie, de Pierre Lafon sur les « rafales »).

Mesures de construction

Il s'agit ici de valider un lien (pas nécessairement binaire) : regroupement d'unités, regroupement de textes, en particuliers.

- *factorisation* : rapport du volume de données enregistrées en utilisant le lien, au volume de données enregistrées sans utiliser le lien (le lien doit être « rentable », à savoir regrouper des éléments qui sinon répètent essentiellement les mêmes structures).
  ex. : indice de Jaccard, indice de Dice.

---

[12] « Un thème peut être latent ou saillant, selon que des constituants sont épars ou regroupés. » (Rastier 1995a, §I.a, p. 228)
(Dupuy 1993, p. 105) cite Charles MULLER (*Dépouillements statistiques et lexicométrie*, 1984 [?]) :
« Daniel Dugast a proposé de distinguer les vocables qui constituent la 'trame' du texte et ceux qui constituent des 'motifs' ; terminologie empruntée aux arts picturaux, si l'on peut dire. La trame, ce serait ce qui apparaît (aléatoirement) un peu partout dans le texte, mots-outils ou non ; les motifs, ce sont alors les vocables qui sont chargés par l'auteur, à un certain moment du texte, et pendant une partie assez brève, de jouer un rôle thématique. »
Outre cette opposition *trame* vs *motifs*, Dupuy utilise aussi une opposition *relief* vs *plan*, qu'il lie aux contrastes entre temps verbaux des propositions d'une même phrase (Dupuy 1993, pp. 291-292).
Les deux propriétés, de répartition ou de regroupement des occurrences, sont complémentaires dans la réalisation de la textualité. Lorsqu'une unité apparaît de façon régulière au long d'un texte, elle concourt à sa *cohérence* et à sa *cohésion* ; une répartition irrégulière en ferait un facteur de *progression* (Dupuy 1993, p. 428).

- *complexité* : nombre de pôles ou dominantes différentes. Il s'agit d'éviter des regroupements volumineux qui jouent sur la souplesse des liens.
  ex. : cardinalité du regroupement formé, sans redoublement (c'est-à-dire que s'il y a plusieurs éléments identiques, ils ne sont comptés qu'une fois).
- *équilibre* : rapport entre les tailles extrêmes des unités regroupées.
  ex. : *min* / *max*
- *cohésion* : taux de recouvrement entre les différentes unités.
  ex. : volume de l'intersection (caractérisations en commun) rapporté au volume de l'élément le plus petit (car c'est celui qui limite le volume possible de l'intersection).
- *significativité* : toutes choses égales par ailleurs, il est préférable d'établir des liens entre des éléments dont le volume d'apparition ou d'usage est le plus conséquent, et d'éviter de bâtir des regroupements sur des cas exceptionnels.
  ex. : volume de l'élément le plus petit.
- *intégrité* : rapport du nombre d'observations où le lien n'est pas validable, mais à peu de choses près, au nombre d'observations où le lien est réalisé. Par exemple, rapport du nombre de cas dans le corpus où une unité descriptive n'est pas réalisée, en raison de l'absence d'une seule de ses composantes, au nombre total d'occurrences de l'unité descriptive.

Mesures d'élection

Il s'agit d'établir l'ensemble d'unités descriptives les plus aptes à caractériser un texte. Les critères suivants correspondent chacun à un mode de lecture du texte.

- *connexité* : complémentarité syntagmatique, à savoir l'ensemble des unités retenues offre globalement une bonne couverture du texte.
  Mode de lecture : lecture régulière (progressive, linéaire, systématique) vs lecture exploratrice (rebonds).
- *cohérence* : complémentarité paradigmatique, à savoir l'ensemble des unités retenues garde toutes les principales thématiques concernées dans le texte.
  Mode de lecture : lecture exhaustive (complète) vs sélective, filtrante (consultation sur un aspect).
- *dominance* : mise en valeur explicite, indiquée notamment par la présentation du texte et par son rapport à un genre. A un genre correspondent des pratiques de rédaction et de lecture conventionnelles, avec des « points d'attente » particuliers : rôle par exemple du début, de la conclusion, etc.
  Mode de lecture : lecture superficielle (survol rapide et synthétique, lecture en diagonale, s'appuyant sur ces éléments dominants) vs lecture détaillée (analytique).
- *saillance* : mise en valeur contextuelle, par exemple importance gagnée par les aspects qui sont originaux dans ce texte, dans le cadre du corpus.
  Mode de lecture : lecture contrastive (*a posteriori*) vs lecture descriptive (*a priori*).
- *dynamique centripète / centrifuge* : caractérise une unité en fonction de ses rapports avec les autres unités du texte : est-elle fédératrice, centrale, proche des autres unités, (elle est alors centripète), ou est-elle marginale, et tournée plutôt vers d'autres textes, élargissant ainsi le domaine du texte (elle est alors centrifuge).
  Mode de lecture : lecture prudente, approfondie vs lecture audacieuse, élargie.

c) Le dispositif des catégories, pour articuler quantitatif et qualitatif

Les unités caractérisant un texte ne sont pas toutes issues de mesures quantitatives ; elles peuvent aussi être qualifiées par des informations, d'ordre qualitatif. Or très souvent, dans les moteurs de recherche en particulier, ces deux types d'information sont confondus, et l'information qualitative est rabattue sur une expression quantitative (directement utilisable dans les calculs). Par exemple, le décompte des occurrences d'une unité fournit une fréquence : c'est bien un indicateur quantitatif, et

ce serait perdre une partie de son information par exemple que de le traduire par « unité peu fréquente », qui peut-être ne permettrait pas de savoir s'il est apparu une seule fois, ou deux ou trois fois (or on a vu que cela peut être très significatif). Inversement, si l'utilisateur veut indiquer qu'un des mots-clés de sa requête doit davantage orienter la recherche que les autres, il est très discutable de lui demander d'indiquer un « poids » ou de le répéter. En effet, il n'a aucun moyen de contrôler l'impact de tel chiffre plutôt que tel autre, alors que l'indication qu'il veut transmettre s'exprime clairement en termes qualitatifs : tel mot-clé est important pour l'usager. C'est au moteur de recherche d'utiliser au mieux cette information (en la traduisant éventuellement par un poids, adapté au fonctionnement interne du système), non pas à l'usager d'indiquer des valeurs dont il ne maîtrise pas la signification effective.

Les catégories sont le mode de représentation de l'information qualitative associée aux unités caractérisantes (c'est-à-dire, les unités descriptives dans le contexte d'un texte). Chaque information peut en fait être présentée sous une forme quantitative ou qualitative, l'une seule des deux formes étant sa forme originale (source), l'autre étant une traduction éventuelle pour des calculs ou pour une interprétation synthétique. L'essentiel, c'est de respecter la nature de l'information originale, pour l'intégrer de la façon la plus adaptée dans le traitement.

Les catégories sont donc une innovation importante par rapport à une représentation purement quantitative (vecteurs pondérés). Dans le cadre de l'application de diffusion ciblée, elles permettent :

- structurellement, d'associer à chaque unité une part d'information qualitative. Jusqu'à présent, l'information entrant en ligne de compte était purement quantitative (nombre d'occurrences et pondération). Or par exemple, si un destinataire averti complète ou modifie manuellement son profil, les informations ajoutées et les retouches ne doivent pas avoir le même statut, pour le système, que les informations calculées automatiquement ; elles sont explicitement validées, et plus stables au fil des mises à jour (il ne convient pas de les effacer brutalement à chaque renouvellement de la base des profils).
- algorithmiquement, de donner une sémantique différente aux différentes unités d'un profil, se traduisant par des traitements distincts, en particulier pour le calcul des similarités document - profil. Les unités prennent alors un rôle précis pour la détermination des rapprochements. Par exemple, l'absence d'une unité peut être pénalisante ou non, la présence d'une unité obligatoire ou à l'inverse éliminatoire, la cooccurrence de certaines unités significative ou redondante, etc.
- du point de vue de l'ergonomie, de donner à l'utilisateur ou à l'administrateur du système un mode d'action propre et bien défini. Sans les catégories, l'ajustement d'un profil ne pouvait se faire que de façon brutale (suppression d'unités), ou aveugle et inélégante (modification « à vue de nez » des valeurs des poids). L'affectation d'une unité à l'une ou l'autre des catégories –ou le changement de catégorie– évite de déformer et de perdre l'information quantitative réelle, et permet de la combiner avec une information supplémentaire gérable par le système.

Première application des catégories : termes transverses et termes de métier

Ce paragraphe présente un travail qui est exposé de façon complète dans (Bommier, Lemesle 1995).

Les unités n'ont pas toutes le même pouvoir de caractérisation dans le contexte d'EDF ; des algorithmes ont été mis au point pour repérer des classes de termes induites par l'organisation des activités (structuration en Groupes, Départements, Services), et capter ainsi une part de l'implicite lié au contexte d'émission des textes :

- certaines unités sont des termes généraux (ex.: `AGENT EDF`, `PROTOTYPE`, `ESSAI`,...) qui peuvent être utilisés pour décrire la plupart des secteurs d'activité de la DER. Ils sont appelés *termes transverses* et distingués par une catégorie propre[13].

  Une méthode issue de la théorie de l'information construit, pour chaque corpus (muni d'une structure sous-jacente, par exemple celle des départements de la DER), une liste des termes transverses.

[13] La catégorie associée aux termes transverses a par exemple pour effet de prévenir un rapprochement qui ne se ferait que sur des termes de ce type, au détriment de termes plus spécifiques qu'ils occultent.

- certaines unités sont discriminantes (ex.: NEUTRONIQUE) : elles sont spécifiques à un domaine d'activité particulier : ce sont les *termes discriminants*.
- certaines unités sont au cœur de la définition d'une activité particulière, mais sont aussi largement utilisées dans d'autres contextes. Il s'agit alors de *termes définitoires*. Par exemple, DOCUMENTATION est un terme définitoire (et non discriminant) pour le Département SID (*Systèmes d'Information et de Documentation*) ; car dans toute la DER il pourra être question de la documentation d'un logiciel.

  Ces deux dernières classes sont étiquetées avec la même catégorie, celle des *termes de métier*, bien descriptifs d'une activité particulière. Pour les identifier, a été mise au point une méthode qui repose sur le calcul de probabilités simples d'apparition d'un terme dans un contexte local ou global. Elle fait usage de cinq indicateurs complémentaires (cinq mesures parmi celles recensées au paragraphe précédent), combinés sous forme de neuf critères permettant de repérer les termes de métier. Un système de pondération nuance le résultat ; il intègre notamment l'évolution de l'usage des termes sur plusieurs années.

2. Utilisation

a) Le profil d'approche

Observations préalables

Dans le modèle de l'espace vectoriel, le texte est représenté par un vecteur qui indique les unités qui le caractérisent et avec quels poids. Les confrontations texte - texte sont alors typiquement effectuées par un produit scalaire ou un cosinus.

Les unités de pondération très faible ont alors un rôle mitigé. Leur avantage est éventuellement d'apporter un contexte, qui supplée des décalages terminologiques pour les unités principales. Autrement dit, deux textes dont les mots-clés principaux ne se correspondraient pas pour des raisons de synonymie par exemple, peuvent néanmoins être mis en relation via des mots-clés secondaires, qui mettent en valeur la similarité des contextes thématiques des deux textes. Mais les unités de pondération très faible ont aussi l'inconvénient d'alourdir la représentation et de multiplier les calculs : elles sont nombreuses, pour une efficacité réelle souvent négligeable.

Le nombre d'unités à considérer est un compromis. Si ce nombre est très grand, il entrave les performances du système, voire exige des ressources trop grandes. Si l'on considère trop peu d'unités, alors la caractérisation des textes (et profils) est trop grossière, et ne permet pas de bien les distinguer les uns des autres, et les résultats de calculs de proximité manquent de nuances.

Définition d'un nouveau mode de confrontation des représentations des textes

Seules les unités les plus caractérisantes de chaque texte, et notamment celles de plus haut niveau, servent dans un premier temps au calcul des similarités texte-texte : c'est le *profil d'approche* du texte. De fait, le rôle contextuel des unités secondaires est moins sensible, car les unités de haut niveau sont les unités qui ont déjà un ancrage contextuel (dans notre modèle, ce sont des Communautés par exemple). Le profil d'approche est également défini en tenant compte de la complémentarité des unités retenues, pour assurer une représentation de l'ensemble du texte.

Au besoin, et dans un second temps seulement, la représentation est élargie aux unités annexes, en utilisant le *profil complet*. Le but de cette démarche est d'affiner la caractérisation du rapprochement, notamment pour différencier plusieurs textes rapprochés par des unités semblables, lors du premier calcul. Une autre raison de recourir aux profils complets est d'élargir l'exploration du corpus, lorsque les résultats sur profils d'approche sont jugés insuffisants. Il est raisonnable de réserver cette recherche élargie à un second temps de la recherche, car elle est moins rentable : le profil d'approche repère l'essentiel des rapprochements intéressants, l'extension générale aux profils complets encombre les résultats de multiples rapprochements de faible intérêt.

Cette démarche correspond à une démarche claire et sensée : on commence par aller d'abord à l'essentiel, pour ensuite approfondir les points qui le demanderaient. Elle contribue dans le même temps à l'efficacité opérationnelle de l'application.

b) Interactions entre les unités

Observations préalables

Les interactions entre mots clés sont très mal gérées, si l'on s'en tient à la combinaison linéaire d'un produit scalaire (ou d'un cosinus). Un mot-clé isolé peut à lui seul susciter un rapprochement fort. A l'autre extrême, une accumulation de mots-clés mineurs (en fait abusivement élevés au rang de mot-clé) et sans relation les uns avec les autres peut également être à l'origine d'un rapprochement.

La prise en compte systématique de toutes les unités est pénalisant, et pas uniquement sur le plan de la rapidité des calculs. Si un texte (ou un profil, pour la diffusion ciblée) comporte plusieurs thématiques, cette représentation pénalise les rapprochements sur un seul thème. Effectivement, tout rapprochement partiel n'est pas nécessairement valide ou pertinent : mais un rapprochement partiel qui reste dans le cadre d'une des thématiques centrales devrait ressortir.

Prise en compte assouplie et contrôlée

Les unités descriptives définies dans cette thèse représentent déjà des indications plus contextuelles que les mots clés : une unité descriptive peut rendre compte de l'association sémantiquement significative de plusieurs mots-clés. Ceci est mis à profit pour réduire le risque de dérive par rapprochement sur un mot-clé isolé décontextualisé.

Les unités du profils d'approche sont munies de l'un des trois rôles suivants, d'impact décroissant :

- *centralité* : l'unité fait parti des points essentiels de la représentation du texte (par exemple, c'est un thème développé tout au long du texte, ou mis en relief par un surlignage) ; alors, l'absence de l'unité parmi les unités en communs motivant un rapprochement est pénalisante ; en revanche, sa présence peut suffire à justifier la sélection d'un rapprochement.
- *fiabilité* : l'unité est caractéristique de la thématique du texte, elle la situe bien ; sa présence parmi les termes en commun dans un texte peut suffire à justifier la sélection d'un rapprochement ; mais à la différence du cas précédent, l'absence de l'unité n'a pas d'effet négatif.
- *significativité* : l'unité apporte des éléments sémantiquement significatifs, qui valorisent un rapprochement déjà assuré par une unité dotée de fiabilité ou de centralité. Sa présence est un plus par rapport à un lien dont la validité est déjà garantie ; son absence n'est pas pénalisante.

Les rapprochements ne sont plus uniquement sanctionnés sur un seul score numérique, trop synthétique. Ils suivent des règles qui modulent les combinaisons intéressantes d'unités. La confrontation de deux textes s'opère comme suit :

- mesure de la *force* du lien, qui est une fonction de l'ensemble des unités descriptives en commun entre les profils d'approche des deux textes (fusion des unités caractérisantes).
  Si la force du lien est non nulle, alors deux autres indicateurs permettent de relativiser la valeur du rapprochement par rapport aux autres rapprochements trouvés sur le corpus :
- mesure de l'*assurance* du lien, qui est une fonction de deux ensembles, celui des unités de centralité et de fiabilité du premier texte présentes dans le second, et celui des unités de centralité et de fiabilité du second texte présentes dans le premier.
- mesure de la *pénalité* associée au lien, qui est également une fonction de deux ensembles, celui des unités de centralité du premier texte non présentes dans le second, et celui des unités de centralité du second texte non présentes dans le premier.

Cette manière de faire évite de fondre indistinctement ces trois mesures, qui ont chacune leur signification et leur rôle. Elles contribuent différemment d'abord au repérage de l'existence d'un lien, puis à son évaluation. C'est en cela une innovation majeure dans la conception des mesures de similarité texte - texte.