Pr. Chiali FZ
Université d'Oran

9. Numérisation, Lemmatisation et traitement lexicométrique

Cette recherche anime le débat qui oppose les chercheurs quant à l'utilisation de la lemmatisation comme mode d'organisation lexico-métrique.

Plan

1. De la numérisation à la lemmatisation

Le texte établi, à partir d'une saisie scanner, est ensuite soumis à des procédures d'analyses informatisées qui exigent un codage explicite, l'essentiel de ce codage est la lemmatisation. Il s'agit de mettre en place les conditions d'une exploration où la statistique prendra part, dans cette perspective et pour réduire le poids des hapax ou des classes trop faiblement représentées, le regroupement de formes s'impose.
Pour l'essentiel, les procédures de lemmatisation ne s'écartent guère de la norme et de la terminologie lexicologique adoptée dans les travaux de l'INALF, notamment du Trésor de la Langue Française (T. L. F). Le principe de lemmatisation est de regrouper le plus de vocables sous une seule forme[27] : (ainsi, tous les verbes conjugués seront remis à l'infinitif, tous les pluriels seront remis au singulier si celui-ci apparaît au moins une fois dans le texte, les participes eux, seront admis comme des formes verbales etc.)
- Nous distinguons les noms des adjectifs (par retour au contexte, le plus souvent), de même que la distinction substantif/verbe. Ainsi, par exemple "abruti" dans la séquence N24/40 du roman Nedjma de Kateb Yacine :
Ex : *Le voyageur n'est plus qu'un abruti, en guenilles ; il attend l'été pour (...)*
est devenue :
Le voyageur ne être plus que un *abruti*, en guenilles ; il attendre l'été

De "*abruti*" dans la séquence N 80/12 :

Ex : *Moi j'ai pas sommeil, y en a qui sont abrutis par le soleil ; Lakhdar et, (...)*
est lemmatisé en :
moi je avoir pas sommeil, y en avoir qui être *abrutir* par le soleil ; Lakhdar et moi (...)

- Les verbes sont ramenés à l'infinitif, les participes passés et les participes présents à valeur d'adjectif sont considérés comme des formes verbales et sont de ce fait ramenés à l'infinitif.

[27] Les formes fléchies sont chargées en mémoire, puis répertoriées dans un fichier Excel. La démarche suivante est de faire correspondre pour chaque forme fléchie le lemme correspondant. L'opération est réalisée manuellement.

- Nous ne conserverons pas la marque du genre ; tous les féminins seront ramenés au masculin. Exemple : beau, beaux, belle, belles sous le seul lemme *beau*[28]. Mais un certain nombre de substantifs féminins sont conservés y compris ceux qui relèvent d'un couplage morphologique avec un correspondant masculin : *Maîtresse* par exemple par opposition à *Maître*.
Ex_1 : *La mère de Kamel avait été à Constantine l'une des rares maîtresses que Si Mokhtar eût gardées plusieurs années,(...)* (Nedp.96)
Ex_2 : *Voyant son maître écrasé par la hargne, la bonne se réfugie à l'étable. Il y a des années qu'elle a perdu le sommeil, comme les vaches qu'elle passe ses nuits à inquiéter, de sa fantomatique présence.* (Ned p.14)
Dans ce cas, le critère est purement sémantique. Plusieurs relectures du texte lemmatisé demeurent indispensables pour corriger toutes les erreurs de lemmatisation.
L'objectif de la démarche de lemmatisation est de modéliser une densité textuelle aussi homogène que possible, cette homogénéité est garantie par la constance de notre démarche lemmatisante.

Dans l'exemple de Nedjma de Kateb Yacine, la lemmatisation nous conduit à un texte de 83420 unités (mots) contre 83669 dans le texte original. D'un vocabulaire de plus de 10725 formes différentes, on passe à environ 6688 lemmes. On trouvera dans les résultats, répertoriés sur CD, le vocabulaire établi par recensement des lemmes, avec l'indication de la fréquence. Nous nous garantirons contre toute dérive interprétative en posant le retour au texte (retour systématique au cotexte des occurrences) comme principe méthodologique systématique. La tâche nous incombe d'être tout le temps vigilant aux différentes affinités (en termes de polysémie) de la langue et de la langue littéraire.

2. La lemmatisation : Pour ou Contre ?
Ces choix pour lesquels le chercheur opte, problématisent la pratique même de la lemmatisation *c'est-à-dire sur l'unité d'indexation et de décompte en lexicométrie*. L'opposition entre les tenants de la lemmatisation (Charles Müller) et les partisans du "mot-graphique", comme lieu unique d'incidence des variables descriptivo-textuelles (Maurice Tournier), est vive.
Considérée comme un perfectionnement de la lexicométrie par les puristes, la lemmatisation est un exercice «périlleux» et contre-productif pour les formalistes. Ces derniers estiment que le fait de lemmatiser *c'est aller à l'encontre même de l'approche lexicométrique qui entend déconstruire le plus objectivement possible un texte pour accéder à son sens ».*
Reprenant l'ensemble du débat entre formalistes et lemmatiseurs, Damon Mayaffre propose de renommer la discipline « logométrie ». Aujourd'hui grâce au logiciel Hyperbase, il est possible de traiter des textes bruts et des textes lemmatisés simultanément. Ce produit permet d'accorder ces deux visions de l'analyse des textes afin de fournir des résultats objectifs et linguistiquement fiables.

3. L'outil informatique et lexicométrie
Depuis 1960, la lexicométrie politique a connu un épanouissement considérable et des travaux riches gérés par les institutions de «Lexicométrie et textes politiques» de l'ENS Saint-Cloud » ou par la revue *Mots/Ordinateurs/Textes*. Le développement d'une lexicométrie littéraire initiée par Guiraud (Guiraud, 1954), Charles Muller (Muller, 1967) ou d'Etienne Brunet sur l'ensemble de la littérature française (Brunet,

[28] Le lemme sera noté en italique.

1981) se reverra exprimée par le biais de logiciels et des applications ajustées aux textes étudiés.
Matériellement, par exemple, la disponibilité de textes numérisés de plus en plus nombreux et de bonne qualité éditoriale, sous un format universel XML, non seulement favorise mais réclame une approche automatique et quantitative. Là où les chercheurs étaient arrêtés dans leur premier mouvement par la fastidieuse saisie numérique des textes, ils se trouvent aujourd'hui noyés par des données textuelles informatisées de plus en plus vastes et immédiatement disponibles sur le Web ou ailleurs. Concomitamment, le développement d'outils lexicométriques toujours plus puissants rend possible le traitement de ces macro-corpus textuels. Longtemps limitées aux traitements d'ensembles de 250.000 ou 500.000 occurrences, les capacités des logiciels sont sans cesse repoussées, pour donner à l'outil - nécessaire supplétif à l'œil ou à la mémoire humaine à partir d'un certain volume - toute sa raison d'être.
Bref, la disponibilité et l'abondance des corpus numérisés d'une part, l'amélioration des capacités des logiciels d'autre part sont la condition *sine qua non*, désormais remplie, du redémarrage de la linguistique quantitative assistée par ordinateur.

Très vite, en effet, les linguistes ont souligné la vanité du traitement lexicométrique car celui-ci s'arrêtait à la matérialité graphique des textes. De fait, le «mot», pris dans sa définition la plus restrictive, ne recouvre pas une réalité linguistique opérante pour permettre la compréhension des textes. Ainsi la lexicométrie peut être soupçonnée de permettre, au mieux, une description du contenu matériel «de surface» des textes, et aucunement d'en recouvrer le sens. Au fond, elle serait un gadget coûteux en temps, sans grande pertinence scientifique.
Les progrès récents des logiciels de lemmatisation, articulés à la nouvelle génération des logiciels de lexicométrie aboutissent à une mutation et à une amélioration de nos pratiques statistiques sur les textes: c'est ce que nous appelons le glissement de la lexicométrie originelle vers une logométrie pleine et entière, susceptible de renouveler la discipline. Cette amélioration est d'ores et déjà effective pour le français depuis quelques années grâce au développement d'Hyperbase, souvent présenté dans *l'Astrolabe* (Brunet, 2001 et 2003) et qui traite sans difficulté les sorties du lemmatiseur Cordial : lemmatiseur polyglotte Tree Tagger pour permettre le traitement logométrique de textes anglo-saxons et romans.
Mais que l'on ne s'y trompe pas cependant: le propos n'est pas de renoncer au traitement lexicométrique sur textes bruts, il est de compléter ce traitement par une analyse complémentaire sur textes lemmatisés.
Ce travail voudrait rappeler que le traitement des textes lemmatisés qui ouvre la voie à des analyses grammaticales ou syntaxiques nous semble indispensable, mais qu'il ne peut se faire qu'à condition de garder accès au texte réel, natif, brut, que le locuteur/scripteur a effectivement émis.

4. Du mot au lemme

D'évidence, deux opérations doivent être envisagées pour que l'unité d'indexation et de décompte cesse d'être une unité matérielle, aveugle sémantiquement (le «mot»), pour devenir une unité de sens pertinente linguistiquement (le «lemme»): les dégroupements et les regroupements linguistiques.

La première opération - le dégroupement appelé aussi «désambiguïsation» - consiste dans sa plus simple expression, à séparer les homographes pour les rattacher à leur vocable respectif.
La seconde opération consiste schématiquement à regrouper sous un lemme unique (en français l'infinitif pour les verbes, le masculin singulier pour les noms, etc.), les

formes différentes - classiquement les «flexions» - signifiant la même chose ou se rattachant au même signifiant.

Cependant, il est primordial d'insister sur les ambiguïtés d'un décompte qui peut regrouper le substantif ''parti'' avec la participe passé du verbe « partir ». Selon Dominique Labbé, près d'un tiers de la composition des textes français est homographe et, selon Charles Muller, ce taux varie en fonction des productions mais ne descend jamais au-dessous de 15 %. Refuser la lemmatisation, c'est admettre qu'une bonne partie du traitement quantitatif que l'on voudrait objectif - les chiffres donnent cette impression d'objectivité - compte ensemble torchons et serviettes, au motif qu'ils revêtent la même apparence graphique. Le discrédit est particulièrement important car, là où certaines pratiques se contentent de l'intuition pour analyser les textes, le traitement quantitatif aspire à la froide objectivité: il ne saurait donc se faire sur des unités impertinentes linguistiquement.

Par ailleurs, toujours dans le cadre des *dégroupements*, mais au-delà des homographes, il convient de dire un mot des formes contractées car elles représentent une surface non négligeable des textes. Effectivement la contraction, qui répond avant tout à une logique économique de la langue, concerne, par définition, des unités très souvent utilisées comme «du» ou «au»; unités fréquentes qui pèsent donc dans les décomptes). Ainsi «du» doit être dégroupé en «de le». Non par luxe bien sûr, mais par souci d'équité car sinon une discrimination linguistique artificielle entre la formule féminine «de la» et la formule masculine «du» serait arbitrairement créée.
Dès lors, les conséquences mathématiques seraient automatiques: on trouverait beaucoup plus de «la» dans les textes que de «le» (dont une partie serait fondue dans «du»), nous laissant imaginer à une féminité du discours. De la même manière, il convient de dégrouper «au(x)» en «à le(s)». Et ainsi de suite.

La pertinence des *regroupements* est moins directement évidente et surtout moins innocente comme nous le démontrerons plus bas. Mais elle reste difficilement contestable dans certaines de ses tâches élémentaires. «Bel» et «beau» ou «nouvel» et «nouveau» doivent être regroupés dans l'index des formes sous une seule entrée. Ils n'ont pas à être comptés distinctement, sinon leur poids se trouverait divisé par rapport au poids des autres adjectifs tel «laid» ou «ancien» qui ont - du point de vue quantitatif - l'avantage d'avoir une forme unique. «Je» et «j'» ont-ils besoin d'être distingués lorsque les autres pronoms personnels ne souffrent pas de diamorphisme?

Autrement dit, souvent deux lexies sont strictement synonymes et seules des contraintes morpho-linguistiques expliquent leur diversité graphique: dans ces conditions comment une approche qui prétend traiter objectivement le texte pour en retrouver le sens peut-elle justifier de les considérer séparément, au risque de diluer leur poids ou leur fréquence dans le corpus par rapport à d'autres mots concurrents qui ne connaissent pas de dispersion graphique?
Par ailleurs, le *regroupement* de «clin d'œil» en un seul vocable n'a guère besoin d'être justifié puisque «clin» n'est pas même une entrée des dictionnaires. Compter «clin» indépendamment reviendrait à considérer quelque chose qui n'existe pas sémantiquement. Et ce constat caricatural sur «clin d'œil» doit être généralisé à toutes les lexies composées («chemin de fer», «pomme de terre», «président de la république») qui ne constituent qu'un seul vocable et dont il ne serait pas seulement absurde mais dangereux sémantiquement de compter les différentes composantes.

La lemmatisation des textes est révolutionnaire et que, devenue accessible à tous, elle ouvre la voie à de riches pratiques qui nous paraissent susceptibles de refondre la discipline. Ramener une forme fléchie à son lemme implique une reconnaissance d'informations linguistiques essentielles. Ramener «dira» au verbe dire, c'est savoir que cette graphie dans le texte est un verbe et que celui-ci prend cette forme à *la troisième personne du futur*. Dès lors, les critères quantitatifs - dont nous n'avons plus besoin de souligner le raffinement aujourd'hui - peuvent s'appliquer à ces informations linguistiques essentielles et nous pourrons analyser l'usage statistique des verbes *versus* les noms, mais aussi l'usage des personnes verbales (ici la troisième du singulier), mais encore du temps dans un corpus donné.
En bref, la lemmatisation permet d'une part de décompter des unités à la pertinence linguistique plus avérée (les lemmes qui renvoient à des vocables), et d'autre part de compléter le traitement purement lexical traditionnel par un traitement statistique d'autres régularités linguistiques tels les codes grammaticaux, les modes, les temps, le genre, le nombre, etc. Dans ces conditions, il apparaît aujourd'hui difficile de se priver d'un tel enrichissement.
Les lemmatiseurs-étiqueteurs affichent, dans leurs fonctions de base (désambiguïsation des homographes, regroupement flexionnel des noms ou des verbes, reconnaissance des catégories grammaticales élémentaires), un taux de réussite total pour ces fonctions de bases quelques erreurs subsistent mais aucune susceptible de fausser le traitement statistique que nous opérons pas la suite; les vérifications manuelles restent utiles mais non plus obligatoires.

5. Logométrie

Evidemment cette conclusion apparaîtrait angélique si elle n'était pas techniquement possible et d'ores et déjà effective. Hyperbase s'applique aujourd'hui dans toutes ses fonctionnalités à décliner ses traitements simultanément sur le texte brut et sur le texte lemmatisé. L'ensemble de l'ergonomie du logiciel a été conçu à cet effet (lire nécessairement ici même, Brunet, 2003). Dès l'activation du bouton «Lecture» qui permet la consultation du texte du corpus, la fenêtre de l'écran est divisée et l'on lira en parallèle le texte brut et le texte lemmatisé
La distance intertextuelle, les recherches thématiques, les AFC, les analyses arborées seront calculées ou construites parallèlement selon les formes graphiques, les lemmes ou les catégories grammaticales. Les fonctions documentaires traditionnelles en lexicométrie, comme la recherche de concordances, de contextes, de co-occurrences, etc., fonctionneront aussi, indifféremment, sur le texte natif et sur le texte transformé et étiqueté.
A vrai dire, aujourd'hui, il n'est plus question de faire un choix exclusif entre formes et lemmes mais de savoir comment organiser techniquement les différentes entrées dans le corpus pour combiner les différents points de vue sur le texte. Hyperbase a choisi une étude parallèle en juxtaposant les analyses et en adaptant, à moindre frais, la statistique *lexicale* traditionnelle à une statistique *grammaticale* et *syntaxique*.

Conclusion

La *Logométrie* est un ensemble de traitements documentaires et statistiques du texte qui ne s'interdit rien pour tout s'autoriser; qui dépasse le traitement des formes graphiques sans les exclure ou les oublier; qui analyse les lemmes ou les structures grammaticales sans délaisser le texte natif auquel nous sommes toujours renvoyés.
C'est finalement un traitement automatique global du texte dans toutes ses dimensions: graphiques, lemmatisées, grammaticalisées. L'analyse ainsi portera sur toutes les unités linguistiques, de la lettre aux isotopies, les mots, les lemmes, les codes grammaticaux, les bi-codes ou les enchaînements syntaxiques.

Un texte est un tout dont le fonctionnement linguistique est complexe; son traitement ne saurait se borner à un seul point de vue.

Bibliographie

- BRUNET (Étienne), Logiciel *Hyperbase* (version 1.6), Institut National de la Langue Française, 1993.
- DUPUY (Jean-Philppe), *Analyse sémiolinguistique informatisée d'Un texte littéraire : Méthodes et application aux nouvelles orientales de Marguerite Yourcenar*, Thèse de Doctorat, Université de Besançon, 1993.
- JUILLARD (Michel), «Études quantitatives des champs sémantiques et morpho-sémantiques dans une œuvre littéraire*.» In la recherche française par ordinateur*, Genéve-Paris, Slatkine-Champion, 1985.
- MASSONIE (Jean-Philippe), *Analyse informatisée des textes*, université de Besançon, 1990.
- MASSONIE (J-P), «Q-occurrences libres» *In BRUNET*, *Méthodes quantitatives et informatiques dans l'étude des textes*. 611-623 - Slatkine-Champion, 1986.
- MÜLLER (Charles), *Initiation aux méthodes de la statistique linguistique*, Paris Hachette, 1977.
- PECHEUX (Michel), *Analyse automatique du discours*, Paris, Dunod, 1969.
- VUILLEMIN (Alain), *Informatique et littérature, Paris,* Champion-Slatkine *(1950-1990),* 1990.