Lexicométrie sur corpus étiquetés

Bénédicte Pincemin

Lexicométrie sur corpus étiquetés

Bénédicte Pincemin[1]

[1]CNRS, LLI – Univ. Paris 13 – Av. J-B Clément – 93430 Villetaneuse – France

in Le poids des mots, *Actes des 7es journées internationales d'analyse statistique des données textuelles (JADT 2004), G. Purnelle et al. (éds), vol. II, pp. 865-873.*

Abstract

Tagged corpus are now widely available, and are of great interest for textual and linguistic studies. Some lexicometric softwares have new versions to handle such corpus, but these don't give complete satisfaction yet. However, a clear and powerful model of text for lexicometric procedures has been formalized, as a string of *positions* ; in each position one or several *types* are instanciated, from one or several sets of types, such as a set of spellings, or a set of lemmas, or a set of grammatical codes.
As regards the types definition, the way these kinds of linguistic information are recorded (the *record axes*) should not be confused with the views one can wish for a lexicometric analysis (the *analysis axes*). Actually, record axes are often irrelevant analysis axes. As regards the string of positions, some positions may be removed for the purposes of the analysis, so as to define the appropriate *background* retained from the *text.* Then the analysis can also be focussed on a given *pattern*, standing out against the background. We finally propose means to complete the results' display. These are naturally expressed and organized according to the analysis axis, but the introduction of views from some other axes may clarify, adjust or enrich their interpretation.

Résumé

Devant la disponibilité et l'intérêt des corpus étiquetés, l'adaptation des logiciels de lexicométrie n'est pas encore pleinement satisfaisante. A cependant été explicité un modèle lexicométrique du texte, comme suite de *positions* en chacune desquelles s'instancie un *type*, et ce éventuellement pour plusieurs familles de types (graphies, lemmes, codes grammaticaux par exemple).
Il convient d'abord de pouvoir redéfinir des *dimensions d'analyse* fixant les types considérés, qui ne sont pas directement la reprise des dimensions d'enregistrement des informations dans l'étiquetage, celles-ci étant généralement non pertinentes si elles sont utilisées telles quelles. Quant aux positions, il est intéressant de pouvoir en masquer certaines (*filtre fond / texte*) puis de focaliser les calculs sur un motif donné (*sélection forme / fond*). Enfin, bien que les résultats doivent rester présentés selon la dimension d'analyse qui les structure, nous proposons des manières de leur associer des éclairages selon d'autres dimensions, pour clarifier, nuancer ou enrichir l'interprétation.

Mots-clés : statistique textuelle, linguistique de corpus, lemmatisation, étiquetage morphosyntaxique, interface, Weblex, Hyperbase.

1. Avancées et limites des réalisations actuelles

Les corpus étiquetés se développent et deviennent les terrains d'analyse privilégiés pour les études textuelles et linguistiques. Le cas le plus fréquent d'étiquetage, que nous développerons ici, est celui de l'étiquetage morphosyntaxique, réalisable automatiquement sur de grands volumes de textes. Un tel étiquetage segmente le texte en mots (sur des critères plus linguistiques que le découpage par repérage de caractères délimiteurs) et associe typiquement à chaque mot son lemme, sa catégorie morphosyntaxique et ses informations de flexion.

Les logiciels de lexicométrie, initialement conçus pour traiter des corpus segmentés en graphies, se sont ouverts aux corpus étiquetés selon des voies relativement différentes.

L'approche la plus simple consiste... à ne rien changer. Les informations apportées par l'étiquetage peuvent être elles-mêmes considérées comme des « mots », et l'on peut faire des calculs lexicométriques sur le corpus écrit comme une succession de lemmes, ou encore écrit comme une succession de codes grammaticaux, ou tout simplement écrit comme la succession de ses mots. Les limites de cette approche sont évidentes : les calculs se placent dans l'une ou l'autre de ces vues, sans qu'aucun lien ne puisse être fait entre les occurrences se correspondant en fait d'une vue à l'autre. Les informations apportées par l'étiquetage sont perçues isolément, et chaque point de vue sur le corpus n'est réalisable que par le calcul (lourd) d'une nouvelle base, sans lien outillé avec les précédentes.

La seconde approche, illustrée par exemple par Hyperbase (Brunet, 2001 et 2002), consiste à intégrer ces multiples vues dans une même base. Le corpus est consultable avec son étiquetage complet. Pour les calculs généraux (vocabulaire présent avec ses fréquences, spécificités), l'affichage des résultats peut être donné en parallèle selon différentes vues intéressantes à considérer ensemble, par exemple graphies et lemmes. Mais en définitive, chaque calcul ne porte jamais que sur une vue du corpus, sans possibilité de croiser les différentes informations.

La troisième approche, développée par Weblex (Heiden, 2002), redéfinit la modélisation du corpus en lexicométrie pour la généraliser. Finalement, ce que l'on considère, c'est une suite de positions ; jusqu'à présent, à une position était associée une information (la graphie) ; avec les corpus étiquetés, c'est une série de propriétés qui peuvent être attachées à chaque position. Un objet d'étude peut alors être sélectionné par une équation combinant différentes propriétés : par exemple, on peut s'intéresser aux adverbes terminés en *-ment*, ou encore aux adjectifs qualificatifs suivant un verbe *être* non auxiliaire à la troisième personne. Ce modèle s'inspire du moteur de recherche mis au point par Christ (1994), et Weblex s'appuie sur cet outil. Les calculs lexicométriques apportés par Weblex restent cependant menés selon une vue à l'exclusion des autres (cf. paramètre *attribut d'occurrence implicite*) — à l'exception du calcul d'index. Une fois l'objet d'étude sélectionné, on considère toutes ses occurrences sous forme de graphies, ou de lemmes, ou de codes grammaticaux, tant pour les calculs que pour l'affichage.

Dans ces contextes, la pratique de la lexicométrie sur corpus étiquetés, enrichissante par la nouveauté des points de vue qu'elle permet, révèle néanmoins des limites importantes.

Tout d'abord, les analyses sur les codes grammaticaux sont peu exploitables à cause de la dispersion des informations, chaque propriété morphosyntaxique étant émiettée à proportion des autres propriétés auxquelles elle est associée au fil des occurrences. Il faudrait plutôt pouvoir considérer certaines propriétés et neutraliser certaines autres, en fonction des objectifs de l'analyse. Ainsi, Habert et Salem (1995), dans une analyse lexicométrique sur corpus étiqueté, en viennent à retravailler l'étiquetage soumis aux calculs, et concluent : « l'analyse des décomptes portant sur l'utilisation d'un système de catégories moins détaillé donne une image que nous avons jugé plus intéressante ».

De même, les autres objets d'étude (lemmes, graphies) ne sont pas si clairs ni satisfaisants qu'il paraît de prime abord (cf. § 2.2).

Enfin, le choix d'une dimension d'analyse (la graphie, le lemme, le code grammatical) s'avère être aussi celui de l'information affichée. Or pourquoi ne pas imaginer une concordance effectuée (et triée) sur les lemmes accompagnés des informations complètes de

flexion, mais affichant également, pour plus de commodité, la graphie correspondante ? Ou encore, on pourrait sans extravagance vouloir le relevé du vocabulaire d'un corpus lemme par lemme, tout en détaillant pour chaque lemme les formes fléchies attestées dans le corpus.

Les approches actuelles semblent donc loin d'avoir définitivement résolu la prise en compte des corpus étiquetés dans les calculs lexicométriques. Montrant des perspectives extrêmement riches (comparaison des calculs selon différentes vues, définition d'objets d'étude par de multiples informations linguistiques complémentaires, généralisation du modèle lexicométrique) tout en frustrant l'utilisateur par les limites qui persistent, les réalisations actuelles sont motivantes et invitent à poursuivre la réflexion si bien lancée.

2. Modélisation

2.1. Le modèle lexicométrique du texte

Les calculs lexicométriques considèrent les textes comme une suite de positions, chacune de ces positions réalisant une occurrence d'une unité type. L'étiquetage s'intègre très bien à ce modèle en considérant que chaque position est multidimensionnelle, c'est-à-dire porteuse de plusieurs informations (graphie, lemme, code grammatical, etc.) (Heiden, 2002) qui sont elles mêmes autant d'occurrences renvoyant à des familles de types différentes (les types sont alors l'ensemble des graphies différentes, ou l'ensemble des lemmes, ou l'ensemble des codes grammaticaux). Cette modélisation ne rend évidemment pas compte de toute la réalité linguistique (où les délimitations d'unités ne sont pas toujours franches ni identiques d'un niveau d'analyse à un autre), mais elle offre néanmoins un terrain d'analyse fructueux.

2.2. Dimensions de codage, dimensions élémentaires et dimensions d'analyse

Ainsi, le texte n'est plus seulement une suite de mots, mais il est aussi, et même en même temps, une suite de lemmes ou une suite de codes grammaticaux. Cependant, chacune de ces dimensions prise indépendamment n'est pas toujours un bon objet d'étude au plan linguistique, et il faut donc se donner la possibilité de redéfinir des objets d'étude pertinents à partir des informations à disposition sans être limité à les reprendre telles quelles.

Prenons d'abord le cas des graphies : considérer les graphies isolément est parfois fait pour retrouver le cadre de travail d'avant l'étiquetage, pour évaluer, par comparaison, les changements apportés par l'étiquetage. Or les outils d'analyse linguistique peuvent être capables de reconnaître des locutions et des mots composés, ou du moins segmentent le texte sur des critères non purement typographiques (utilisation d'un dictionnaire, de listes de mots, de grammaires). Ne peuvent donc pleinement coexister une dimension de découpage par reconnaissance de caractères délimiteurs (Lebart et Salem, 1994, § 2.1.2) et des dimensions de lemmes et de codes grammaticaux. Du coup, une analyse qui serait sur les graphies seules perd beaucoup de son intérêt, puisque cette vue n'est qu'une approximation du traitement par découpage simple et correspond moins bien à une réalité linguistique claire que les graphies désambiguïsées. Les adeptes de la non-lemmatisation[1] se retrouveront mieux en effet avec les graphies désambiguïsées, qui combinent la forme fléchie des mots et le code grammatical (il serait équivalent de donner le lemme, puisque toutes les informations de flexion sont précisées dans le code grammatical). Cette dimension d'analyse distingue tout ce qui

[1] Les arguments scientifiques développés par exemple dans (Geffroy et al., 1974) ne sont pas périmés par la disponibilité actuelle de corpus étiquetés.

morphosyntaxiquement peut être distingué, sans opérer de regroupements *a priori*, permettant de ne pas masquer des divergences de comportement statistique entre les différentes flexions d'un même lemme.

Les adeptes de la lemmatisation quant à eux ne devraient pas choisir de travailler sur l'étiquette « lemme », mais sur les lemmes désambiguïsés, c'est-à-dire où l'information lexicale de la graphie du lemme est associée à l'information grammaticale de sa catégorie[2]. Car bien souvent, ce que l'on appelle le lemme n'est en fait que la graphie du lemme. Ainsi, on ne s'intéresse pas aux occurrences de *devoir* ou de *boucher*, mais à celles de *devoir (nom)* et de *devoir (verbe)*, de *boucher (nom)* et de *boucher (verbe)*.

Enfin, les codes grammaticaux ont rarement un intérêt à être exploités tels quels et pour eux-mêmes. Ils combinent de multiples informations qu'il sera généralement plus clair d'étudier séparément, voire de neutraliser quand les distinctions opérées ne semblent pas pertinentes pour l'étude. Habert et Salem (1995) choisissent par exemple, pour le dépouillement des réponses à une question ouverte (corpus *enfants*), de ne considérer les flexions que pour les verbes conjugués. Pour le français, les étiquetages morphosyntaxiques fournissent des informations de deux natures : catégorie et flexion. La catégorie précise les « parties du discours » (nom, verbe, etc.) affinées en sous-catégories quelquefois par des « traits » transversaux (indéfini, interrogatif, etc.) qu'il est intéressant de pouvoir considérer indépendamment d'une partie du discours particulière. Quant à l'information de flexion, elle n'est pas toujours présente (un certain nombre de catégories sont invariables) ; elle couvre les variations en genre, nombre, personne, et la conjugaison (mode et temps) pour les verbes.

Cette particularité du code grammatical, qui se décompose en fait en plusieurs informations structurées, pourrait théoriquement également advenir dans les dimensions plus lexicales de la graphie ou du lemme : pourquoi pas coder en effet des composants lexicaux comme les radicaux, préfixes et suffixes, les morphèmes, etc.

Il faudrait donc finalement distinguer :

- les *dimensions de codage* : typiquement graphie fléchie, graphie du lemme, code groupant catégorie et flexion. Elles sont héritées des usages en matière d'étiquetage. Il n'est pas certain qu'il faille en rendre compte pour l'utilisateur, puisque ce n'est pas *a priori* pertinent pour l'analyse lexicométrique.

- les *dimensions élémentaires* : par exemple morphèmes (si disponibles, ou sinon graphies des formes fléchies et graphies des lemmes), catégories, sous-catégories et traits, flexion en genre, nombre, personne et conjugaison (modes et temps). Elles sont souplement combinables pour la définition de dimensions d'analyse.

- des *dimensions d'analyse* : principalement graphie fléchie (ou du lemme) + catégorie + flexion, graphie du lemme + catégorie, tout ou partie des informations de catégorie ou/et de flexion éventuellement combinées avec des informations lexicales (lemme). A la différence des dimensions de codage et des dimensions élémentaires, en nombre limité fixé par le format d'enregistrement, les dimensions d'analyse ne sont pas énumérées mais construites par l'utilisateur en fonction de ses objectifs. Certaines (comme la graphie ou le lemme, désambiguïsés) sont évidemment plus classiques que d'autres et pourraient être proposées de façon prédéfinie.

[2] La procédure d'intégration de corpus étiqueté de Weblex définit une telle propriété d'occurrence.

- et éventuellement des *dimensions d'affichage* : elles se construisent comme des dimensions d'analyse, et ne se distinguent que par leur rôle (le calcul s'appuie sur la dimension d'analyse choisie, et les dimensions d'affichage ne font qu'apporter des éclairages complémentaires pour l'interprétation des résultats, cf. partie 2.3.3). L'affichage peut motiver la construction de dimensions qui lui sont propres (peu propices aux calculs mais intéressantes pour l'exploitation des résultats), telles que la graphie fléchie seule qui, en contexte (par exemple dans une concordance), est plus lisible et n'est pas ambiguë.

2.3. Les trois moments de l'analyse lexicométrique

Etant donné un corpus découpé en unités linguistiques étiquetées, le modèle lexicométrique du texte laisse donc trois paramètres à préciser. Deux sont nécessaires pour le lancement d'un calcul : d'une part, le domaine d'action du calcul, autrement dit l'ensemble des positions, des unités linguistiques, qui forment l'image du corpus pour le calcul (c'est ce que nous appellerons plus loin la délimitation du *fond* dans le *texte*) ; d'autre part, ce qui permet, en chaque position considérée, de se rapporter à un type dans un paradigme de types réalisables dans le corpus. Un troisième paramètre apparaît pour la présentation des résultats : si ceux-ci correspondent à une dimension donnée, l'affichage d'autres dimensions peut faciliter leur lecture ou éclairer leur interprétation.

2.3.1. Les étapes de sélection : définition du fond dans le texte, focalisation sur des formes se détachant sur le fond

Dans les logiciels actuels, le premier paramètre annoncé ci-dessus est peu disponible. Il n'apparaît peut-être que dans des options de découpage lors de la compilation d'un corpus. Les élagages et filtrages proposés ne délimitent pas le *fond* dans le *texte*, mais contribuent plutôt à sélectionner les *formes* sur le *fond.*

En effet, un corpus soumis à une analyse lexicométrique comporte en fait trois strates incluses l'une dans l'autre. La première est le *texte*, qui ancre l'ensemble des positions disponibles, se matérialisant typiquement par la suite complète des mots, des ponctuations, des signes typographiques. Les deux autres strates sont le *fond* sur lequel se profilent les *formes*[3]. Le fond sert au décompte des positions qui donne le contexte de référence dans les calculs (taille du corpus et des parties, domaine de décompte des fréquences). La strate des formes est l'objet d'étude, sélectionné parmi les éléments du fond. Dans les affichages d'occurrences en contexte, il est souhaitable en général d'afficher le texte complet (ne serait-ce que pour des questions de lisibilité), *a minima* le fond et les formes qui s'y détachent. Des effets typographiques devraient alors permettre de repérer ce qui relève de chacun des niveaux texte / fond / forme, par exemple le texte en plus petites polices ou en italiques, les formes en gras.

D'où deux filtres de sélection sur un corpus : l'un, souvent négatif, pour définir le *fond* à partir du *texte* ; l'autre, habituellement positif, et quelquefois doublé d'un discret filtre négatif d'« élagage » (classiquement mots grammaticaux, mots de moins de trois lettres, hapax, seuil sur le score statistique), pour désigner les *formes* à reconnaître sur le *fond*, auxquelles on va s'intéresser. Le premier filtre opère en amont des calculs, en influant sur la définition du fond

[3] Dans cette communication, nous utilisons le mot « graphie » pour désigner les formes fléchies, afin de réserver le mot « forme » à cette idée que se donner un objet d'étude en lexicométrie c'est finalement étudier le contraste de formes qui se détachent sur un fond.

qui sert de référence aux calculs il peut modifier, de façon « invisible », les scores associés aux formes. Alors que le second filtre opère en aval des calculs : il sert à montrer les résultats sur un sous-ensemble choisi par l'utilisateur d'un calcul qui par ailleurs aurait souvent pu être mené sur l'ensemble du fond sans autre incidence que de grossir la taille du calcul.

Préciser la forme que l'on souhaite étudier sur le fond est facultatif pour les calculs sélectifs, à savoir pour lesquels un score statistique est calculé et un seuil borne l'étendue des résultats. Faute de forme particulière indiquée, le calcul s'applique alors à tous les types de la dimension d'analyse choisie ayant des occurrences dans le fond, puis le seuil limite l'affichage aux cas les plus remarquables. Pour d'autres procédures, telles que la recherche de contextes ou la construction de concordances, il n'y a pas de score statistique calculé : la sélection forme / fond est alors incontournable, et est d'ailleurs beaucoup plus spontanément abordée (on fait la concordance de tel mot, ou on recherche les contextes de tel lemme).

Si la forme est un motif complexe (sur plusieurs positions), elle peut être soumise en argument aux calculs de scores statistiques (tels que spécificités)[4] dans la mesure où l'on accepte de sortir du modèle statistique initial, et de faire l'approximation que, étant données les tailles du texte et des parties en nombre de positions, la forme peut être considérée comme ponctuelle.

Le travail sur un fond différent du texte n'est pas non plus anodin du point de vue des modèles statistiques. Dans certains cas, on pourra considérer que le fond reste très proche du texte et en reste une bonne approximation. Dans d'autres cas, il faudra plutôt pouvoir convenir, pour lancer les calculs statistiques, que la distribution des types sur le fond accepte les mêmes modèles et lois statistiques que ceux sur lesquels reposent les procédures utilisées.

2.3.2. Définition de la granularité (rapport occurrences / type) par le choix d'une dimension d'analyse

Le deuxième paramètre qui intervient dans les calculs consiste à définir, par une combinaison de dimensions élémentaires[5], la dimension d'analyse qui fixe les types auxquels rapporter les occurrences (cf. § 2.2) : par exemple, le tableau 1 ci-après montre, pour un extrait de phrase, les types reconnus selon différentes dimensions d'analyse. En particulier, c'est selon la dimension d'analyse que s'opèrent les tris. (Notons que lorsque l'on considère des formes « longues » (sur plusieurs positions), chaque position peut se voir attribuer une dimension d'analyse propre. Un calcul n'est donc pas toujours paramétré par une seule dimension d'analyse.)

Concrètement, si l'on considère les verbes (en tant que fond), on pourrait choisir de ne s'intéresser qu'aux temps verbaux (indépendamment des personnes), et étudier leur

[4] Les calculs sur les suites de mots se pratiquent de fait déjà dans les logiciels lexicométriques.

[5] On indique un paradigme en donnant simplement les dimensions de variation pertinentes. Une interface graphique pour ce faire pourrait s'inspirer notamment du masque de saisie conçu pour Hyperbase (Brunet, 2002, figure 9 : *Le choix d'un code grammatical*). Une autre voie pour définir une dimension d'analyse consisterait à définir autant de filtres en cascade que de types : si *condition 1* alors *type 1* sinon si *condition 2* alors *type 2*, etc. sinon *type reste* (hétérogène). Cette voie complémentaire est sans doute moins centrale que la combinaison de dimensions élémentaires, qui par nature sont ajustées à l'articulation linguistique des informations et construisent des paradigmes homogènes.

répartition au fil du texte ; ou encore, on pourrait centrer une analyse sur les personnes, sans les disperser sur les catégories qui les portent (possessifs, verbes, pronoms personnels).

	Dimensions d'analyse		
Extrait de phrase :	graphies désambiguïsées	catégories et sous-catégories grammaticales	flexion en nombre
Le	Le + Da-ms-d	Da (déterminant article)	s
deuxième	deuxième + Ao-ms	A (adjectif qualificatif)	s
paramètre	paramètre + Ncms	Nc (nom commun)	s
qui	qui + Pr-ms--	Pr (pronom relatif)	s
intervient	intervient + Vmip3s-	Vm (verbe principal)	s
dans	dans + Sp	S (préposition)	
les	les + Da-mp-d	Da (déterminant article)	p
calculs	calculs + Ncmp	Nc (nom commun)	p
Nombre de types :	8	6	2 + 1[6]

Tableau 1 : Exemples de types reconnus selon différentes dimensions d'analyse.

Dans les logiciels actuels, ce paramètre est encore souvent réduit à la sélection d'une dimension de codage : par exemple, le paramètre *attribut d'occurrence implicite* dans Weblex, qui rend déjà de grands services. Mais on n'a guère le choix de la granularité, qui est maximale : typiquement, si l'on travaille sur l'étiquetage morphosyntaxique, c'est chaque étiquette complète qui est constituée en type, avec une dispersion évidente et gênante des résultats. Deux amorces de définition souple de la granularité existent cependant. D'une part, pour le calcul d'*index* de Weblex, on peut « composer l'index avec les champs » de son choix, ce qui revient à sélectionner les dimensions actives pour définir les types. D'autre part, dans les *listes* d'Hyperbase, on peut grouper ponctuellement des items successifs ; ce nouvel objet peut être soumis au calcul des spécificités (histogramme) (et aux statistiques sur tableau, telles que analyse factorielle ou arborée). On a des procédés comparables de sélection groupée dans d'autres logiciels, par expression CQP pour les rafales dans Weblex, et par expression régulière dans les *types généralisés* de Lexico3.

2.3.3. Affichages multidimensionnels : dimensions complémentaires et projections de résultats

Le troisième paramètre, relatif à l'affichage, permet d'indiquer une ou plusieurs autres dimensions complémentaires à la dimension d'analyse. Il intervient de façon différente selon le type de résultats produit.

Le résultat peut être des *occurrences en contexte* : extraction de contextes, concordances, mais aussi surlignage en contexte des motifs sélectionnés par un calcul avec seuil (segments

[6] Certaines dimensions d'analyse produisent des valeurs indéterminées pour certaines positions, par exemple ici les prépositions (et les conjonctions) n'ont pas de marque de nombre. On peut considérer que ces occurrences sont à considérer de façon groupée sous un type supplémentaire que l'on pourrait appeler *indéterminé.* Cependant il s'agit d'un type *a priori* hétérogène, et on préfèrera éviter autant que possible la formation de tels types par un filtrage *texte* / *fond* qui écarte de l'analyse les positions régulièrement non décrites par la dimension d'analyse.

répétés, spécificités, rafales,...). L'affichage par défaut pourrait être le texte entier (non réduit au fond), avec indication, aux positions sélectionnées, des types reconnus, suivis des affichages complémentaires demandés. Il reste souhaitable (hors éventuelles considérations techniques de débit et de volume) qu'en tout point le texte soit cliquable, avec affichage complet des informations associées à la position considérée.

Le résultat peut autrement se présenter sous forme de *listes de types*. Dans ce cas, les affichages complémentaires peuvent se présenter comme une liste des valeurs prises par les occurrences de chacune des positions de réalisation de chaque type, pour chacune des dimensions souhaitées. Le choix d'un affichage complémentaire devrait donc préciser l'ordre souhaité pour l'énumération des attestations, qui pourrait être classiquement alphabétique (plus généralement canonique[7]) ou par fréquence décroissante.

Si les formes considérées s'étendent sur plusieurs positions (syntagmes, segments répétés, etc.), les affichages complémentaires seraient sans doute à donner position par position, sans préciser à ce niveau toute la combinatoire effectivement réalisée, pour ne pas surcharger les résultats. Pour en savoir plus, il est toujours possible de relancer un calcul en changeant de dimensions d'analyse.

Des procédures de projection / comparaison pourraient être disponibles, notamment pour visualiser la portée des choix qui ont une incidence sur le calcul : sélection texte / fond (qui fixe le « corpus de référence »), ou dimension d'analyse (par exemple lemmatisation ou non). L'utilisateur pourrait ainsi placer côte à côte plusieurs listes de résultats. Le clic sur un item d'une liste sélectionnerait dans les autres listes les items qui ont des occurrences communes dans le texte ou la partie. Un double clic sélectionnerait une liste entière ; le même principe de projection sur les autres listes permettrait de repérer les items originaux des autres listes, ceux qui n'ont pas de correspondants dans la liste initiale.

3. Conclusion

Face à la disponibilité et à l'intérêt des corpus étiqueté, l'adaptation des logiciels de lexicométrie conduit à généraliser la modélisation lexicométrique du corpus en « positions » et « dimensions » attachées à chacune de ces positions. Cette communication propose quelques éléments pour prolonger et développer une telle modélisation. La notion de *dimension d'analyse* est introduite pour donner accès à des objets d'étude non prédéfinis par les habitudes de codage et à la granularité réellement adaptée aux problématiques de recherche. La détermination des éléments linguistiques soumis au calcul lexicométrique (les *formes* à profiler sur un *fond*, éventuellement différent du *texte* complet) serait alors opérable par la définition de motifs, combinant des éléments correspondant à des positions unitaires dans le corpus. A chaque élément serait associé *(i)* son repérage (critères de sélection pour reconnaître les positions qui réalisent cet élément), *(ii)* sa dimension d'analyse (de quel type, dans quelle liste de types, est-il une occurrence), *(iii)* son affichage (outre la dimension d'analyse choisie en *(ii)*, qui détermine les calculs et articule la présentation des résultats, d'autres dimensions peuvent être demandées pour apporter un éclairage complémentaire).

[7] Nous appelons *ordre canonique* un ordre d'énumération méthodique et mnémonique, associé aux dimensions élémentaires et aux dimensions d'analyse lors de leur construction. L'ordre canonique de codes grammaticaux peut être différent de l'ordre alphabétique : on voudra par exemple placer *singulier* avant *pluriel*.

Malgré son appareillage théorique central, l'exposé garde un caractère expérimental et prospectif. Il vise à défricher un terrain de réflexion et d'expérimentation sans prétendre fixer un cadre de référence stable. En effet, à ce stade de notre cheminement, ce sont la conception et éventuellement la réalisation de développements concrets qui seront maintenant les mieux à même de faire le tri entre les propositions fructueuses et celles impraticables, comme de préciser ce qui n'a été que trop vite défini. Un travail de conception d'interface est déjà amorcé, non moins passionnant que cette première élaboration du modèle théorique.

Références

Brunet E. (2001). *Hyperbase, logiciel documentaire et statistique pour la création et l'exploitation de bases hypertextuelles. Manuel de référence.* Version 5.2 (mai 2001). Laboratoire Bases, corpus et langage, UFR Lettres, Université de Nice.

Brunet E. (2002). Le lemme comme on l'aime. Proc. of JADT'2002, pp. 221-232.

Christ O. (1994). A Modular and Flexible Architecture for an Integrated Corpus Query System. Proc. of COMPLEX'94 (3rd Conf. on Computational Lexicography and Text Research), pp. 23-32.

Geffroy A., Lafon P. and Tournier M. (1974). L'indexation minimale. Plaidoyer pour une non-lemmatisation. ENS de Saint-Cloud. Communication au *Colloque sur l'Analyse des corpus linguistiques : Problèmes et méthodes de l'indexation minimale*, Strasbourg, 21-23 mai 1973.

Habert B. and Salem A. (1995). L'utilisation de catégorisations multiples pour l'analyse quantitative de données textuelles. *Traitement automatique des langues*, 36 (1-2):249-275.

Heiden S. (2002). *Weblex. Manuel Utilisateur*. Version 4.1 (janvier 2002), Laboratoire ICAR, UMR 5191, ENS Lyon.

Lafon P. (1984). *Dépouillements et statistiques en lexicométrie*. Slatkine-Champion.

Lamalle C., Martinez W., Fleury S., Salem A., Fracchiolla B., Kuncova A., Maisondieu A. (2003). *Lexico 3, Outils de statistique textuelle. Manuel d'utilisation.* Version 3.41 (février 2003), Laboratoire SYLED – CLA2T, Université de la Sorbonne nouvelle Paris 3.

Lebart L. and Salem A. (1994). *Statistique textuelle*. Dunod.

Rajman M., Lecomte J. and Paroubek P. (1997). Format de description lexicale pour le français – Partie 2 : description morpho-syntaxique. Technical report GRACE, GTR-3-2.1.