Mots dans le texte, mots hors du texte: réflexions méthodologiques sur quelques index et concordances appliqués à des œuvres françaises, italiennes ou espagnoles

par

Suzanne Hanon

«... Duggan: *A Concordance of the Chanson de Roland*: Ein willkommenes Arbeitsinstrument für Linguisten, Literarhistoriker und Textkritiker, das in seiner mechanischen Stupidität nur von einem Computer gemacht werden kann ...» (K. B., Besprechungen, *Zeitschrift für Romanische Philologie*, 86 (1970), p. 676).

1. Introduction

La publication récente de travaux éminents portant sur de larges corpus en philologie romane amène certains à se poser la même question: pourquoi les auteurs de ces travaux ne se sont-ils pas servis des inventaires exhaustifs déjà établis pour prouver définitivement certains faits ou pour étayer les théories avancées? De tels inventaires existent et portent le nom de concordances ou d'index. Ils permettent la consultation rapide, le simple contrôle ou l'étude approfondie de n'importe quel problème philologique et, partant, d'une série d'autres disciplines. Malheureusement, ces inventaires sont mal connus, parce que parfois mal distribués, ou ils sont d'accès difficile à divers niveaux et restent donc lettre morte pour la plupart des chercheurs. Le présent article a pour but de montrer au potentiel usager romaniste ce qu'il peut attendre de ces inventaires, comment ils sont élaborés, quels types de dépouillements existent, quelles sont leurs limites et quelles manipulations les textes ont parfois subies pour se prêter à ces remaniements.

2. Le texte

Si nous acceptons que le texte, quel qu'il soit (et nous n'entrerons pas ici dans la discussion de ses limites ou de sa clôture), est l'objet d'étude pri-

vilégié du linguiste ou du chercheur en littérature, alors ce même texte réarrangé de façon pratique pour pouvoir être lu, relu, analysé, copié, critiqué, passé sous la loupe du syntacticien, du lexicographe, de l'historien de la culture ou de la littérature, ce même texte nous est rendu, «docile», «apprivoisé» en quelque sorte, maniable, sous forme de concordance. Nous pouvons aussi établir que le texte en question n'est que le support matériel d'idées: en général, les linguistes s'attachent à étudier le niveau «matériel» ou formel du texte, tandis que les littéraires dépassent rapidement cette barrière pour essayer de pénétrer la pensée de l'auteur, ses intentions, etc. Ceci nous ramène en définitive à une discussion des unitées du texte, les mots, ces unités mal définies et continuellement remises en question par les chercheurs et qui sont, bon gré mal gré, l'objectif ou le truchement de l'objectif des chercheurs. Les concordances et les index sont des inventaires de mots, et on peut dire que, contrairement aux dictionnaires, vocabulaires, glossaires et lexiques (qui sont aussi des inventaires de mots mais des mots hors du texte, des mots de 'langue'), ces inventaires définissent une fois pour toutes l'ensemble des mots d'un texte, donc aussi l'ensemble des emplois de ces mots, leur combinatoire, mais aussi l'ensemble des sens de ces mots. L'avantage premier est donc ici la limitation naturelle d'un corpus et, à l'intérieur de celui-ci, d'un ensemble de traits supposés avoir quelque chose de commun: les mots dans le texte. Les autres inventaires comme les dictionnaires ne sont, la plupart du temps, que des approximations, puisque le corpus n'est «fini» qu'en théorie; les vocabulaires ou glossaires d'auteurs sont d'habitude non-exhaustifs, c'est-à-dire qu'ils ne répètent pas une information déjà donnée, donc à leur sens superflue, et en cela, ces inventaires manipulent donc l'information donnée puisqu'ils estompent, pour mieux en mettre d'autres en relief, des parties du texte qui auraient pu, par leur présence même, constituer une information.

3. Historique des concordances

Ces remarques préliminaires étant faites, il faut souligner que les concordances ont joui depuis des temps immémoriaux d'un certain crédit, et disons même d'une certaine crédibilité dans les études de la Bible. R. Brackenier (1972), qui retrace dans son article, les origines bibliques des concordances, mentionne, comme première compilation, entre autres les *Concordantiae breves*, répertoire de la *Vulgate* compilé par Hughes de Saint-Cher et ses confrères dominicains et datant de 1230 environ. Ces concordances des textes bibliques ont fait leur chemin depuis et on peut citer plus près de nous les

importants travaux d'Ellison: *Nelson's Complete Concordance of the Revised Standard Version of the Bible*, la *Concordance des Quatre Evangiles* de Claire Bompois etc., pour ne rien dire des travaux consacrés aux Pères de l'Eglise qui font l'objet d'analyses très poussées à l'Université de Louvain (cf. P. Tombeur (1969) et J. Hammesse (1972–75)). Les auteurs classiques ont aussi eu les faveurs des concordances, probablement parce qu'il s'agit là de langues mortes, et ils ont parfois bénéficié de trouvailles «artisanales» du genre des fiches inventées par Lane Cooper, lesquelles permettent la compilation relativement rapide de matériaux linguistiques volumineux (par exemple, la concordance des œuvres d'Horace réalisée «manuellement» en un an avec 18 collaborateurs (cf. S. M. Lamb et L. Gould (1964) et IBM Processing Application «Literary Data Processing» 1971). Comme preuve du succès de ce genre de travaux, le «Repertorium lateinischer Wörterverzeichnisse und Speziallexika» de Paul Rowald mentionne près de 150 inventaires de cette sorte déjà en 1914. Citons également les travaux du Père R. Busa sur les manuscrits de la Mer morte en collaboration avec P. Tassman (cf. bibliographie).

Les travaux sur les auteurs modernes datent pour la plupart du XX^{e} siècle et tirent parti de l'essor des techniques mécanographiques (ex. le Laboratoire de Besançon) ou des études sur ordinateur. Mais, encore de nos jours, il n'est pas impossible de mener à bien, à l'instar des moines des *scriptoria*, de façon artisanale (dite «manuelle»), des inventaires exhaustifs de corpus étendus, témoin les efforts fructueux, à Cornell University, du Professeur Cooper, qui travailla avec des équipes d'étudiants et de «femmes au foyer», main-d'œuvre de tous temps exploitée, à ses fameuses concordances sur fiches. Témoin aussi les travaux sur Dante, *A Concordance to the Divine Comedy*, éditée par la Dante Society of America, dont la plupart des membres ont été mis à contribution: plus de cent personnes (préface p. vii–ix) pour mener à bien cette énorme entreprise. Témoin encore le monument à la mémoire de Lope de Vega, le *Vocabulario Completo de Lope de Vega*, compilé par Carlos Fernández Gómez (Real Academia Española, Madrid 1971, vol. I–III), monument auquel l'auteur avoue avoir sacrifié «varios años de esfuerzos continuados, con jornadas laborales a veces de hasta doce horas, cuyo número exacto que conozco no indico por no incurrir en pedantería ...» p. IV).

4. Dépouillements sur ordinateurs

Ces derniers travaux montrent bien que des inventaires plus ou moins exhaustifs requièrent un facteur temps énorme (donc le sacrifice d'une bonne

partie de la vie d'un chercheur) ou une multiplication des effectifs de travail (grand nombre de collaborateurs) ou encore de moyens (mécanisations diverses ou automatisation), ce qui a toujours été le lot et le prix payé par les auteurs de dictionnaires. Ladislav Zgusta souligne cet aspect non négligeable de la tâche du lexicographe dans son manuel de lexicographie (chap. viii). Avant de passer plus avant, qu'il nous soit permis ici, et une fois pour toutes, de redresser quelques torts et de mettre fin à quelques mythes, qui vont généralement de pair avec l'idée qu'on se fait des concordances:

1. Les travaux de lexicographie sont, à notre avis, des travaux savants à part entière et à égalité avec d'autres travaux dits scientifiques.
2. L'élaboration dite «artisanale» des inventaires de vocabulaire n'a rien d'artisanal ni de manuel dans les acceptions négatives que prennent parfois ces deux termes: ce sont aussi, à part entière, des travaux intellectuels, souvent très avancés, très nuancés, mais qui ont de l'artisan le manque de mécanisation, d'automatisation. On devrait plutôt les appeler travaux non-automatisés.
3. Les ordinateurs ne sont ni plus ni moins stupides que les linguistes. Ce sont des appareils visant à la solution rapide de problèmes divers, dont la répétition d'un très grand nombre d'opérations de classements, de comptages, etc. Ils reproduisent, en général, exactement ce qu'on leur a enjoint de faire, ni plus ni moins, et en ce sens, ils ne diffèrent pas beaucoup d'autres appareils à fonction de répétitions diverses comme les réveille-matin, les mixeurs à soupe, les tourne-disques automatiques. Il faudra donc s'attendre à de bons ou à de mauvais résultats selon ce qu'on aura soi-même investi comme effort.

Les ordinateurs ont un gros avantage sur les chercheurs:

a) ils peuvent manier d'énormes quantités de matériaux à la fois;
b) ils permettent des contrôles nombreux de l'exhaustivité des matériaux;
c) ils ne «comprennent» pas ce qu'ils font, ce sont des exécutants qui ne se mêlent en rien aux problèmes posés, ils n'interfèrent pas;
d) du fait que ce ne sont pas des spécialistes, il faut leur inculquer patiemment toutes les notions nécessaires, ce qui exige de la part du chercheur qu'il repense toutes ses définitions, ou qu'il fasse un choix dans la terminologie embrouillée qui lui est impartie en tant que linguiste, spécialiste de la littérature, etc. (en effet, qu'est-ce qu'un mot, une forme, une proposition subordonnée, une métaphore, une connotation?).

Il n'en reste pas moins vrai que la primauté revient au chercheur, même s'il se sert de l'expédient de l'ordinateur. La plupart des concordances citées ci-après ont été élaborées à la machine.

5. Concordances: définitions

La définition la plus générale d'une *concordance*, c'est un réarrangement du texte original pour faire 'concorder', c'est-à-dire mettre en parallèle tous les mots du texte, le modèle le plus traditionnel consistant à réécrire par ordre alphabétique chaque mot du texte original au milieu d'une ligne et à faire en sorte que le contexte naturel du mot choisi soit réparti à gauche et à droite du mot mis ainsi en vedette. Chaque ligne du texte original se trouve ainsi réécrite en plusieurs exemplaires pour permettre à chaque mot d'en être la *vedette* (ce qu'on appelle aussi *mot-vedette* ou *mot-clef*). Il est d'usage de donner à chaque ligne du texte une référence numérique, qui renvoie à l'endroit précis où le mot de cette ligne se trouve: page, ligne ou vers, chapitre ou autres subdivisions naturelles du texte original.

Par opposition avec la concordance proprement dite, l'*index* des mots (*index verborum*) est une concordance où l'on a fait l'économie des contextes en se limitant à donner la référence numérique seule.

L'usage montre malheureusement que, comme pour d'autres termes linguistiques, on a souvent confondu ces deux mots, concordance et index, en employant l'un pour l'autre. La plupart des dictionnaires, même spécialisés, ne donnent qu'une idée très vague de ce genre de travaux. Les manuels de lexicographie sont également parcimonieux sur ce chapitre (J. et Cl. Dubois (1971), J. Rey-Debove (1970) et L. Zgusta (1971)) et ne citent qu'à de très rares exceptions les concordances (Dubois (1971), mais aussi Dubois *et al.* (1974)). Et pourtant ce sont bien des documents lexicographiques, puisqu'ils nous renseignent sur la présence ou l'absence d'un mot, sur sa ou ses valeurs, ses emplois, etc. On pourrait même dire que les concordances sont des dictionnaires particuliers, où les définitions sont remplacées par le contexte naturel qui donne au mot-vedette sa valeur, ou, sans aller si loin, que les concordances sont à mi-chemin entre le texte et le dictionnaire, puisqu'elles respectent le texte de départ mais doivent être lues comme des dictionnaires.

6. Mots: définitions

Mais qu'est-ce qu'un mot? C'est là une question extrêmement pertinente, quoique, dans la pratique, bien difficile à résoudre. La question a en tout cas

fait couler beaucoup d'encre (cf. à ce sujet Togeby (1949), Pottier (1962 et 1966), Muller (1963), Martinet (1966), Wagner (1967–1970), Dubois *et al.* (1973)).

6.1. Mots dans le texte

Si nous reprenons notre texte original, nous pouvons dire que ce texte comprend un nombre *x* défini de mots, c'est-à-dire de formes verbales, de noms, de pronoms, d'adjectifs, de particules diverses. Si le texte contient 138 mots, j'obtiendrai donc une concordance (ou un index) de 138 lignes, me renvoyant chaque fois à une instance de chaque mot dans le texte: un verbe conjugué, un nom, un adverbe, etc. Le mot dans le texte, souvent fléchi, est donc plutôt une *forme* (verbale, adjectivale, nominale, etc., ou une particule); ces formes portent parfois le nom d'*item* (cf. *Des tracts en mai 68*, p. 21). Je peux choisir de compter ces formes ou non. Si je compte le nombre des exemples ou occurrences de *chevaux* dans mon texte, j'obtiendrai une statistique quelconque en mettant *chevaux* en rapport avec d'autres mots ou formes de mon texte, par exemple avec d'autres pluriels, ou avec d'autres mots commençant par *ch-*, ou encore avec d'autres instances comme *cheval* au singulier, etc. En choisissant *cheval* comme *forme canonique*, c'est-à-dire la forme privilégiée prise arbitrairement pour représenter le groupe et en réunissant toutes les occurrences de *cheval* et de *chevaux* sous cette même forme canonique, on opère ce qu'on appelle une *lemmatisation*: on regroupe sous une même *entrée*, ou *lemme* ou *forme canonique*, les formes fléchies d'un même mot. Ce faisant, on introduit un élément arbitraire. Peut-être la forme canonique n'est-elle pas du tout présente dans le texte de départ? On impose donc une nouvelle dimension au texte, on l'idéalise en quelque sorte, on est déjà en train de quitter le fait de parole pour le ranger dans le fait de langue, qui est le propre du dictionnaire. Cette opération, la lemmatisation, peut avoir bien des avantages, mais comme elle se prête à des manipulations diverses, il n'est pas sans intérêt de souligner la part d'arbitraire qu'elle instaure dans le texte vierge. Nous y reviendrons.

6.2. Mots hors du texte

Si nous nous occupons de fréquences, il n'est pas superflu de rappeler qu'un dictionnaire est une liste (en général alphabétique) de mots «hors du texte», qui apparaissent, en tout cas pour ce qui concerne les *entrées* (les titres) ou encore les *adresses* (Quemada, *Les dictionnaires du français moderne 1539–*

1863, 1968, p. 266), une seule fois. Il n'est donc pas très intéressant de faire une statistique de cette liste, car tous ses membres ont la même fréquence, qui est 1. Le dictionnaire est donc un fait de langue. Par contre, dans un texte comprenant des phrases, c'est-à-dire aussi un message à donner, les mots apparaissent avec des fréquences variées, certains même avec des fréquences extrêmement élevées: ce sont pour la plupart des mots courts, prépositions, conjonctions, articles ou pronoms, et d'autres mots encore que l'on est convenu d'appeler *mots-outils* ou *mots vides* («incolores») par opposition aux *mots pleins*. Si nous essayons de faire un compte des fréquences des mots employés dans un texte, nous obtiendrons à coup sûr une liste de mots utilisés un grand nombre de fois (comme *de*, *que*, *le*, *un*, *et*, etc.) et d'autres rarement, peut-être même une seule et unique fois (*hapax*). Dans la série des mots employés plusieurs fois, il faudra distinguer le mot en tant que *type*, c'est-à-dire le représentant du groupe, et le mot comme *token*, c'est-à-dire l'instance de texte, l'occurrence. Exemple: le type *de* est représenté par 686 tokens *de* dans un échantillon de l'*Heptaméron* de Marguerite de Navarre. Cette distinction est extrêmement importante en statistique linguistique, car elle est essentielle dans la division entre dictionnaires (liste de *types*) et concordances (liste de *tokens*).

6.3. Autres considérations sur la notion de mot: le mot graphique

Le découpage de la chaîne du texte pose encore d'autres problèmes pour le moins épineux. La chaîne du texte, prise par exemple dans sa représentation graphique, comporte des «blancs», des signes de ponctuation et des lettres employées seules ou groupées. On peut établir conventionnellement que tout mot est une suite de lettres (suite pouvant se réduire à l'unité) séparée par des blancs ou des signes de ponctuation. Tout en faisant fi du contenu du mot, cette définition est parfaitement acceptable car elle est très favorable aux comptes numériques et aux manipulations sur ordinateur. Elle ne fait d'ailleurs qu'entériner le «sentiment populaire» de ce qu'est un mot. Malgré ce biais de la graphie, qui peut apparaître, dans tout son formalisme, comme une solution de facilité, on enregistre encore, lors de l'analyse, une série de difficultés, qui ne sont pas dues à l'emploi de l'ordinateur mais bel et bien à des circonstances d'ordre taxinomique en grammaire, ou à des syncrétismes de formes s'expliquant parfois sur le plan diachronique ou bien causés par le hasard. Les difficultés d'ordre *taxinomique*, c'est-à-dire de classification grammaticale ou linguistique, apparaissent lors du tri des mots graphiques: *du* représente-t-il un seul mot ou est-ce la réalisation de *de* et *le* (deux mots)?

En d'autres termes, faut-il rattacher *du* au lemme *de* et au lemme *le*? Et, allant plus loin, faut-il voir en *de* un seul ou plusieurs mots (*de* préposition, *de* particule précédant certains infinitifs, *de* article partitif réduit à sa plus simple expression)? Tous les mots contractés posent ce problème (*au*, *aux*, *des*, *lesquels*, etc.), mais c'est aussi le cas des mots élidés: *l'*, *d'*, *m'*, etc. A ce sujet on peut se demander si l'apostrophe est un facteur de rapprochement ou de distanciation, ou les deux à la fois. Que penser aussi de la syntaxe multiple du trait d'union en français (*un je-ne-sais-quoi*, *vient-il*, *week-end*, etc.)? Pas très éloignées des problèmes de classification, on trouve également des difficultés dues à l'*homographie* (deux mots différents ayant la même forme, la même étiquette linguistique) ou à la *polysémie* (une même étiquette présentant des sens différents), qui sont en somme deux faces du même problème, si l'on ne fait pas entrer en ligne de compte des facteurs discriminatoires d'ordre historique ou étymologique. Le relatif *que* est-il homographe du pronom interrogatif *que* ou de la conjonction *que*? Y a-t-il vraiment deux verbes *voler* en synchronie ou bien un seul puisque les deux sens sont issus d'une même origine? La forme *nous* représente-t-elle plusieurs mots puisqu'elle peut être, *mutatis mutandis*, l'équivalent de *il*, *le*, *lui* conjoint et *lui* disjoint? Que faire de *maintenant* adverbe et de *maintenant* participe présent du verbe *maintenir*, etc.? Toutes ces questions et bien d'autres se posent à celui qui collationne des matériaux bruts de mots, et il est difficile d'y répondre de façon univoque. Muller (1963) a bien montré comment des équipes parallèles travaillant sur un même corpus arrivaient à des résultats numériques divergents dus au manque d'homogénéité dans le choix des principes de base. Sur ce sujet on peut aussi consulter Engwall (1974) ainsi que le travail collectif *Des tracts en mai 68* (1975). Le but du présent article n'est pas d'entrer dans ces discussions techniques, qui sont résolues, du moins partiellement, ailleurs; il est pourtant bon de souligner que le comptage des mots graphiques ne va pas sans prise de position sur les nombreuses formes homographes occupant, qui pis est, le dessus de l'échelle des fréquences! Comme il serait absurde de compter pour elles-mêmes des formes ou étiquettes pouvant recouvrir des contenus différents (donc des *mots* différents), la séparation des homographes doit être pratiquée avant toute indication statistique. Cette séparation est le plus souvent élaborée manuellement par la répartition en plusieurs groupes des occurrences à première vue identiques mais dont le contenu peut être dit différent. La séparation des homographes n'est qu'une étape antérieure à la lemmatisation.

6.4. Mots simples / mots composés

Il n'est pas sans intérêt de rappeler que certains «mots» comportent plusieurs «parties», à l'instar de *pomme de terre*, *chemin de fer*, *parce que*, etc. (ce sont les *lexies complexes* de Pottier (1963) et (1966)) et que ce fait n'est pas négligeable dans les dépouillements et surtout dans le comptage des mots.

7. Le contexte

7.1. Contexte naturel

Le contexte d'un mot peut être défini comme le *milieu naturel* dans lequel ce mot se trouve. Le milieu naturel peut être le contexte du syntagme, de la proposition, de la période ou bien de la ligne, du paragraphe où le mot se trouve. En ce qui concerne la poésie ou le théâtre, ce sont plutôt le vers ou la réplique qui peuvent être considérés comme le milieu naturel. Le fait qu'un mot x se trouve dans la phrase (le texte) y peut être quantifié: on peut compter combien de fois le *type* x est représenté dans y, c'est-à-dire le nombre de *tokens* ou d'occurrences de x. Le milieu naturel de x est donc y ou plutôt ce qu'il reste de y quand on en extrait x, c'est-à-dire les autres mots de la phrase y ou encore les *cooccurrents* de x, les mots qui apparaissent avec lui et forment avec lui des sous-groupes ou syntagmes. Si l'on peut étudier numériquement la répartition des occurrences, on peut aussi étudier la répartition des cooccurrences et la fréquence des deux phénomènes ensemble, que l'on nomme la *cofréquence*. Une description détaillée de ce genre d'étude statistique est donnée par le groupe collectif d'auteurs de *Des tracts en mai 68* (1975), également dans A. Geffroy *et al.* (1973). R.-L. Wagner analyse de façon approfondie la signification de ce concept en syntaxe et en statistique (1967–1970), tandis que Jens Rasmussen esquissait déjà en 1967 la portée que cette analyse pourrait avoir en se conjuguant avec les techniques de l'informatique. (Chez ces auteurs, on parle aussi de *collocations*, de *corrélations* et de *cooccurrences*, mais pas toujours nécessairement avec le même sens.) Zellig S. Harris est d'ailleurs un des premiers à avoir jeté les bases de l'étude de l'environnement des morphèmes: A dans un environnement C–D dans «From Morpheme to Utterance» (*Language* 22, 1946, 161–183). D'autres études apparentées ont été faites sur les occurrences et les cooccurrents: ce sont par exemple les *associations paradigmatiques* de Michon et Potdevin (1973), l'analyse des *groupes* dits *binaires* (Gorcy *et al.*, 1970), etc.

Il est donc de la plus haute importance pour l'auteur de la concordance de faire en sorte que le contexte soit arrangé de façon pratique et maniable pour le lecteur. L'utilisateur, de son côté, devra d'abord étudier comment la concordance est présentée, quel est son «format» (pour employer le jargon des informaticiens), avant de procéder à l'analyse du texte. Le mot-clef et son *contexte droit* peuvent présenter un intérêt linguistique ou stylistique important. On peut donc imaginer d'alphabétiser le contexte droit (en tout ou en parties). Le *contexte gauche* peut aussi présenter un certain intérêt (quels sont les verbes qui régissent la conjonction *si*, etc.). Si la concordance en question est complète, tous les contextes de tous les mots pourront être retrouvés du côté droit. Par contre, si la concordance est sélective, une certaine manipulation du contexte gauche peut être souhaitable.

7.2. Le contexte numérique

L'ordre dans lequel apparaissent les occurrences et leurs cooccurrents une fois fixé, il reste encore à indiquer la source d'où provient l'énoncé ou le segment d'énoncé donné. La source est, en général, une référence numérique (page, vers, etc.), mais elle peut également être de nature nominale (titres de poèmes, autres subdivisions du texte, par exemple les «journées» du *Décaméron* (Barbina (1969)). Il appert que, dans la concordance même, la référence numérique au texte-source peut être considérée comme une espèce de contexte. C'est d'ailleurs la seule source donnée dans les *index de mots*. (Cf. par exemple A. E. Creore: *A Word-Index to the Poetic Works of Ronsard*, 1972).

7.3. Contexte imposé ou optimalisé

Jusqu'à présent nous avons parlé uniquement du *contexte naturel* d'un mot. Il est évident que le contexte peut faire l'objet de remaniements divers: le contexte est alors «imposé» ou «optimalisé», c'est-à-dire qu'il est raccourci, allongé, tronqué, etc., pour diverses raisons. On peut vouloir donner la priorité à certains mots sur d'autres en estompant ces derniers: par exemple, accorder la priorité aux mots pleins en estompant une série de mots grammaticaux. Cette méthode est employée dans l'établissement des *groupes binaires* (Gorcy *et al.* (1970)) dans les travaux à la base de l'étude des tracts de mai (*Des tracts* (1975)). D'autres chercheurs estiment comme J. E. G. Dixon (1974) qu'il faut opérer un choix pour donner un meilleur contexte. Ils emploient alors une technique de manipulation avant l'édition de la concordance du texte (*preediting*). Dixon décrit un procédé employé lors de l'élabo-

ration d'une concordance portant sur les œuvres de Rabelais (concordance non publiée), où le contexte naturel serait la période où le mot apparaît (éventuellement la ligne) mais qu'il réduit de façon judicieuse en formant manuellement des «unités de pensée». Ces opérations peuvent être justifiées, mais il faut bien garder à l'esprit que les gains sont obtenus au prix de certaines pertes d'information, par exemple les collocations grammaticales. Un bon exemple de ce type de concordance 'manipulée' est donné par les *Concordanze del Decameron* (1969), où certains contextes sont réduits à des points de suspension pour les rapprocher d'autres contextes.

7.4. Les limites du contexte

Il va de soi que, théoriquement, le contexte maximal d'un mot *x* peut être à la limite considéré comme le texte entier, pris globalement, alors que le contexte minimal peut être envisagé comme le cooccurrent direct à droite ou à gauche du mot-clef, éventuellement les deux à la fois. C'est là une question de définition ou de convention. Certains chercheurs ont ainsi esquissé le concept de *concordance minimum*, c'est-à-dire la séquence linguistique à laquelle le mot appartient (cf. discussion par Quemada de l'article de Mitterand et Petit, *Cahiers de lexicologie*, 1962). On peut, bien sûr, discuter sur la question de savoir si le mot avec son voisin direct a bien une valeur opérationnelle ou s'il faut faire entrer en ligne de compte d'autres cooccurrents. Entre ces extrêmes de contexte maximal et minimal, il s'avère pratique de considérer comme contexte naturel le syntagme auquel appartient le mot en question ou la proposition, éventuellement même la période entière dans laquelle le mot apparaît.

8. Sortes de concordances

Les principes de classification suivis ici ne s'excluent pas les uns les autres; au contraire, une certaine combinatoire se fait parfois jour.

8.1. Concordances complètes vs. concordances abrégées ou sélectives

Si les concordances ne sont qu'un réarrangement du texte de départ, elles fournissent pour chaque mot une ligne concordée: elles sont donc *complètes*. Ce sont les concordances destinées à fournir le maximum de renseignements puisque, restituant le texte-source et étant générales, elles constituent une

étape sûre et exhaustive vers une recherche textuelle quelconque. Parmi les concordances complètes, citons la concordance de la *Chanson de Roland* par Duggan (1969), celle sur *Bécquer* (Ruiz-Fornells (1970)), etc.

Certains textes sont dès le départ tronqués car le chercheur ne s'intéresse qu'à une partie du corpus: seuls le retiennent les mots pleins, ou les mots grammaticaux, ou les métaphores, ou bien les noms propres, ou encore les noms ayant une fréquence supérieure à 5, etc. Ces concordances sont donc *sélectives*, et le texte de départ est abrégé. Parmi celles-ci on trouve diverses réalisations: les concordances sélectives mais où le gros du texte est préservé. C'est le cas de la plus grande partie des concordances éditées et publiées: *La Fontaine* (par Tyler (1974)), *Racine* (Freeman et Batson (1968)), le *Décaméron* (Barbina (1969)), *Garcilaso de la Vega* (Sarmiento (1970)), *Dante* (Wilkins and Bergin (1965)), tous ouvrages mentionnant une liste plus ou moins longue de mots outils exclus: prépositions à haute fréquence, pronoms personnels, formes fléchies de *avoir* et *être*, etc. La composition de ces listes n'est pas constante de concordance à concordance. Cette exclusion d'un sous-ensemble de la population des mots part, en général, d'un souci financier. D'autres concordances sélectives le sont à cause de l'objectif limité d'une étude: concordances grammaticales (Dolores M. Burton (1968)), concordances syntagmatiques (P. Laurette (1974)).

8.2. Concordances machine, concordances brutes, concordances lemmatisées

Par *concordance machine*, on entend généralement le *listing brut* destiné à être remanié et qui constitue la *sortie machine* ou *output*. Par *concordance brute*, on entend parfois une *concordance de formes*, c'est-à-dire où les mots-clefs ne sont pas lemmatisés et où la séparation des homographes n'a pas eu lieu (ex. Duggan, *Chanson de Roland*; Freeman et Batson, *Racine*), etc. Par *concordances lemmatisées*, on suppose que l'éditeur compilateur a fait subir à l'output de l'ordinateur, de sérieuses manipulations grammaticales ou autres; ainsi des concordances de la *Divine Comédie* de Dante (Wilkins et Bergin (1965)), des concordances du laboratoire lexicologique de Liège (*Chrétien de Troyes*, *Blondin de Nesle*, etc.), des travaux de l'université de Gand (*Charroi de Nîmes*, *Villon*, etc.), c'est-à-dire que le mot-clef donné dans la concordance n'apparaît pas nécessairement dans le texte, mais qu'il peut être imposé par l'éditeur. Citons comme exemple *fil*, mot-clef concordance, mais qui n'est représenté par aucune occurrence de cette forme, les 12 occurrences présentes étant *filz* dans le texte du manuscrit Coislin des œuvres de François Villon (Van Deyck et Zwaenepoel (1974), vol. II, p. 204) ou que l'éditeur

impose à la disposition du texte un préarrangement grammatical (*droit* adverbe, *droit* adjectif, *Guillaume d'Angleterre* (Dubois-Stasse *et al.* (1974), p. 118)).

8.2.1. Les problèmes de la lemmatisation

Les problèmes de la lemmatisation (certains emploient le terme *lemmage*) ont déjà été esquissés plus haut. Ils constituent souvent l'optique d'une école, tout comme il y a les adeptes à tout crin des index contre les adeptes des concordances (voir à ce propos P. Grimal (1966), J. J. Duggan (1966), S. Hanon (1973, 1 et 2), Muller (1974)). Certains prônent une lemmatisation grossière, d'autres une lemmatisation fine, nuancée. Diverses approches pour lemmatiser de façon automatique ont été décrites dans la littérature. Les techniques à 100% automatiques butent contre le gros problème de l'homographie. On peut éventuellement fournir à la machine une liste des homographes les plus courants de la langue traitée. Ou bien on peut se servir d'une préindexation ou preediting (à ce propos voir le point 8.6). Pour automatiser au maximum, on peut fournir une «grammaire» de la langue avec une série de tables de désinences à consulter, une liste des mots irréguliers et, éventuellement, une liste d'homographes. Un autre système consiste à lemmatiser manuellement un mini-corpus et à exiger ensuite que l'ordinateur imite cette méthode: on liste alors les mots que l'ordinateur n'arrive pas à analyser. Pour une description des options choisies, cf. Fossier et Zarri (1975). Les procédés employés sont donc plutôt semi-automatiques. Dubois, Dubois-Stasse et Lavis (*Chrétien de Troyes* (1970)) décrivent une méthode de comparaison automatique entre les formes nouvelles et un dictionnaire préalablement établi pour constituer des 'hypothèses de lemmage'.

8.3. Concordances verbales vs. réelles ou notionnelles

Cette distinction étudiée par Brackenier (1972, p. 10) oppose les concordances de mots (verbales) aux concordances de choses ou d'idées (concordances réelles). Les concordances réelles sont toutes des œuvres anciennes; aucune n'intéresse le domaine roman en premier chef. Théoriquement, les concordances réelles ou notionnelles doivent présenter un certain intérêt, par exemple dans les études thématiques d'auteurs: l'idée du gouffre chez Baudelaire, le concept d'amour chez Marguerite de Navarre, etc. Il n'est pas exclu de penser à l'élaboration de tels dépouillements. Ils requerraient l'emploi de dictionnaires de synonymes et une préindexation des périphrases dénotant un même concept.

8.4. Concordances KWIC vs. KWOC

Le mot-clef est présenté dans son contexte (KeyWord In Context = KWIC), ou bien il est employé comme entrée de dictionnaire et se trouve donc isolé typographiquement de son contexte (KeyWord Out of Context = KWOC). Ces deux formats s'excluent en général l'un l'autre, mais certains compilateurs ont judicieusement combiné les deux formats (laboratoire lexicologique de Liège: *Blondel de Nesle*, *Chrétien de Troyes*, etc.). Le format KWIC présente l'avantage de lister systématiquement tous les mots-clefs les uns sous les autres, ce qui fait apparaître des blocs graphiques qui facilitent la perception visuelle avant même qu'il soit question d'étudier le texte. Si les contextes, surtout le contexte droit, sont alphabétisés, ces blocs sont encore plus perceptibles au premier coup d'œil: apparaissent alors les formules (sur ce sujet dans les chansons de geste, cf. Duggan, *Romania*, 1966), les clichés, les groupes syntaxiques fortement cohérents, etc.

Le format KWOC donne une image également très claire, mais la perception visuelle des blocs est neutralisée, de même que le contrôle rapide de certains éléments ou groupes d'éléments est grandement freiné. Le format KWOC est souvent utilisé pour les concordances lemmatisées, mais ce n'est pas là une condition *sine qua non*. On l'emploie aussi quand on veut, de façon facile, regrouper des variantes graphiques (ex. *ben*, *bene* dans la concordance *Canzoniere di Petrarca*, Accademia della Crusca, 1971, p. 199).

Le format KWIC est très favorable aux études des cooccurrences, donc des études linguistiques au sens étroit. Il convient particulièrement bien à l'étude de la prose. Le format KWOC a surtout les faveurs des textes littéraires, du théâtre. La combinaison des deux formats se prête bien au mélange des données statistiques avec le texte de base. Pour se faire une idée de la différence des deux formats, on consultera avec profit deux concordances sur *Les Fleurs du mal* de Baudelaire, dont la version de Besançon (1965) est de format KWOC, tandis que la version de R. T. Cargo (1965) est de format KWIC.

8.5. Concordances et contexte

8.5.1. La nature du contexte

Le contexte peut être brut, tel quel, ou manipulé (Laurette, Dixon, etc.). Le contexte peut être donné

1) dans un ordre quelconque;
2) dans l'ordre chronologique d'apparition dans le texte-source, par chapitres, livres, journées, poèmes;

3) par ordre chronologique de la parution des œuvres d'un auteur;
4) par l'ordre alphabétique des mots qui constituent le contexte, ou
5) par un mélange de ces classifications.

Pour le type KWIC, il est important de réfléchir, avant d'établir la concordance, à l'emploi qu'on fera de cette concordance. C'est là un facteur non négligeable. L'ordre chronologique peut être pertinent dans la recherche stylistique ou thématique des idées d'un auteur, aussi pour marquer l'évolution des idées, des thèmes. Dans les cas où certains passages sont douteux, l'ordre chronologique peut également constituer un précieux document. L'ordre alphabétique est surtout pertinent dans les études des périphrases, de syntagmes, mais aussi pour les expressions toutes faites, qu'elles soient du ressort de la stylistique ou de la grammaire. En ce qui concerne le choix des contextes pour les études documentaires, cf. le point 8.6.3. ci-dessous.

8.5.2. La longueur du contexte

Nous avons déjà fait remarquer que, selon la nature du contexte choisi, on peut donner des contextes naturels ou manipulés de plus ou moins grande longueur. En ce qui concerne les vers ou les répliques d'une pièce de théâtre, le contexte naturel est bien déterminé à l'avance, mais si le mot devant être concordé est un des derniers du vers considéré, la concordance perd de sa valeur justement pour ces instances du texte. M. Spevack discute ce problème pour sa concordance en un volume sur Shakespeare (*Harvard Concordance to Shakespeare*, 1973) et donne des contextes de plus d'un vers pour assurer la bonne compréhension du texte. C'est aussi l'option choisie par Duggan pour la *Chanson de Roland* (1969), ainsi que pour les concordances produites à l'Institut de Lexicologie de Liège (*Chrétien*, *Blondin de Nesle*). Les compilateurs qui ont négligé de prendre en considération cet aspect, en arrivent souvent à donner des contextes très courts, qui se réduisent même à zéro (cf. la concordance de *Bécquer* par Ruiz-Fornells, par exemple). Certains auteurs ont procédé artisanalement à un allongement du contexte de base: par exemple, les auteurs de la concordance du *Charroi de Nîmes* (1970) ont élargi les contextes originaux d'un hémistiche à des groupes plus grands. Ils arrivent malheureusement à des résultats peu systématiques et ne peuvent remédier à des cas comme *cuens Gilebert* pour l'entrée *cuens* (contexte minimum de Quemada). C'est dans un certain sens aussi le cas de la concordance sur les contes et fables de Jean de La Fontaine de Allen Tyler (1974). Tyler a néanmoins choisi le format KWOC, qui se prête moins à l'élargisse-

ment du contexte. Pour assurer la bonne compréhension du texte, on peut conseiller de donner sept à huit mots graphiques de chaque côté du mot-clé, l'idéal étant bien sûr d'aller jusqu'au bout de la proposition (le point-virgule, le point, etc.).

8.6. Concordances et indexation

8.6.1. Indexation: définition

Par indexation nous entendons toute indication marginale venant de l'éditeur de la concordance, pour interpréter le donné linguistique. L'indexation ne veut donc pas dire lemmatisation, mais elle peut être un pas vers une désambiguation du texte. Remarquons tout de suite que la plupart des textes écrits comportent une série de situations ambiguës, dont les homographes ne sont qu'une petite partie. Au niveau de la ponctuation, du choix des lettres majuscules ou minuscules, du découpage en paragraphes, en vers, en laisses, les facteurs typographiques «visuels» comme les blancs, les espaces, etc. ne sont que pauvrement rendus lors de la mise sur ordinateur du texte-source (sur ce sujet, cf. B. Munk Olsen (1968), p. 55 ss.). La plupart des auteurs de concordances ont dû, lors de l'élaboration de leurs dépouillements, trancher une série de problèmes liés au donné linguistique.

Pour les corpus linguistiques imprimés, il va sans dire que les problèmes d'indexation sont relativement minimaux si l'on n'envisage pas de lemmatiser. Il peut cependant être utile d'opérer une division entre noms communs et noms propres, opération qui suppose en général une indexation antérieure à la mise en concordance, puisque certains noms peuvent être homographes même à ce niveau-là (*Pierre*, *pierre*). On peut également trouver souhaitable de distinguer les mots étrangers des mots du texte en question, ou d'éliminer les citations ou autres corps étrangers au texte. Toutes ces opérations présupposent un certain degré d'indexation.

8.6.2. Textes non transmis par l'imprimé

Dans les textes non transmis par l'imprimé, il peut y avoir des raisons encore plus grandes d'indexer. Pour les concordances qui sont élaborées directement sur un manuscrit et non sur une édition de ce manuscrit, l'éditeur doit décider de toute une série de problèmes: quels sont les séparateurs de phrases ou de syntagmes? Où commence un mot? L'éditeur est même parfois amené à redéfinir le concept de mot au niveau de son propre texte, sans compter les problèmes classiques de la résolution des abréviations, des ligatures et autres

problèmes paléographiques. Le philologue informaticien est souvent obligé d'indexer son texte en introduisant, par exemple, une ponctuation opérationnelle (c'est le cas pour H. Naïs et P. Tombeur; cf. Hanon (1973²)).

Pour ce qui est de l'étude des corpus parlés, une foule de problèmes se posent au chercheur, problèmes qui ne peuvent pas être ignorés par l'utilisateur de la concordance, lequel s'exposerait à tirer des conclusions aberrantes. Le problème de la transcription phonétique, par exemple, ne va pas sans aléas (cf. Sankoff-Cedergreen *in* Hanon (1973²)).

8.6.3. L'indexation automatique

En ce qui concerne l'étude des chartes ou autres documents analysés surtout pour leur valeur de sources d'information ou de témoins d'une époque, c'est-à-dire un domaine où l'on est au-dessus des problèmes textuels mais où l'on cherche la réponse à des questions bien définies comme «quel était le souverain régnant à l'époque de la signature de tel ou tel document?», une série de techniques ont été esquissées très récemment par L. Fossier et G. P. Zarri (1975). Un certain précodage multiplié par une optimalisation des contextes (parenthétisation, descripteurs divers) facilite l'élaboration automatique des index-matière, des index onomastiques et toute espèce de recherche de type documentaire. De plus, l'essor qu'ont pris les *techniques conversationnelles*, par lesquelles un dialogue est mené entre le chercheur et l'ordinateur par opposition aux *techniques batch* (c'est-à-dire la résolution par l'ordinateur d'un problème donné suivant un programme défini, mais sans possibilité d'entrer en dialogue), permet beaucoup de flexibilité dans les dépouillements.

8.7. Concordances inverses

Les concordances ou index inverses sont constitués par des *listings ordonnés par la fin du mot-clef* au lieu d'être alphabétisés par le début du mot: ils font faire de sérieuses économies de temps et évitent des recherches fastidieuses qui consistent par exemple à chercher tous les mots se terminant en *-ment* en français (ou un sous-groupe de ces mots). Ce sont en général des dépouillements à l'état machine, mais certains sont publiés, comme le dictionnaire inverse de Juilland (1965) et la concordance sur G. S. Belli (1970). A peu de choses près, ces dictionnaires ou index inverses correspondent à la vieille idée de dictionnaires de rimes, mise à part peut-être la notion d'exhaustivité qu'impliquent les dépouillements dits complets. On voit le parti que l'on

pourrait tirer de telles compilations si elles existaient au moins pour les auteurs connus. Certains de ces dépouillements font appel à des critères supplémentaires de classification, par exemple la transcription phonétique.

9. Les emplois

Après avoir passé en revue les types de concordances et les divers écueils que peut présenter le manque d'homogénéité dans la terminologie et dans la pratique linguistiques, il est bon de souligner les domaines dans lesquels ces ouvrages constituent une aide pour le chercheur.

9.1. Présence ou absence d'un élément

La concordance constitue un *raccourci* à travers le texte lors de la recherche d'un mot précis ou d'un concept. Elle doit représenter à tous les instants un gain de temps, c'est-à-dire qu'elle doit pouvoir rendre superflue la relecture du texte original en entier pour trouver si un élément appartient ou non à l'ensemble. P.ex. le mot *dans* n'est pas employé dans la *Deffence et Illustration* de Joachim du Bellay: le chercheur peut s'en assurer d'un coup d'œil. Si elle est bien faite et bien conçue, la concordance parvient même à rendre le contrôle avec le texte original inutile. C'est là une des fonctions primordiales des dépouillements exhaustifs et, dans ce sens, on peut se servir des concordances comme d'un dictionnaire du texte en question. Mais, à la différence du dictionnaire, la concordance permet le dénombrement du *mot dans le texte*, du mot dans son milieu naturel, et ce dernier trait est important dans toute recherche, qu'elle soit de nature linguistique, littéraire ou autre.

9.2. Combinatoire ou rejet

Mettant en rapport les mots les uns avec les autres, la concordance peut servir à *associer* certains mots à d'autres, à étudier les groupements, les associations naturelles ou les *rejets*. De ce fait, elle permet de se faire une idée de l'emploi d'un mot *x*, donc aussi de son sens, ou en tout cas de l'aire d'emploi de ce mot. A la limite, si la concordance est le seul document disponible qui rende compte d'un texte, d'une œuvre ou d'une époque, on peut ajouter que le sens d'un mot pourra être défini par la somme de ses emplois. C'est d'ailleurs en partant de cette hypothèse que les pères de l'Eglise ont jeté les premières bases de leurs travaux. Les recherches récentes en analyse litté-

raire se basent sur la même idée (Duggan, pour le style en formule des chansons de geste, *Romania*, 1966; Honeycutt, sur les clichés dans les fabliaux, *Romania*, 1975, etc.).

Il est aussi possible d'opérer n'importe quel *relevé statistique* à partir d'une concordance des groupes de deux, trois ou plusieurs autres éléments cooccurrents, mais cet aspect a déjà été mentionné plus haut.

9.3. Répartition

Si la concordance est bien préparée, elle permet l'étude de l'homogénéité du texte, donnant ainsi accès à la *répartition* des mots à travers le texte, l'œuvre, etc. Muller (1967) a bien montré pour l'œuvre de Corneille la répartition et l'évolution du vocabulaire chez cet auteur. On peut se demander, à partir d'une concordance à entrées alphabétiques ordonnées chronologiquement suivant les années de parution des différentes œuvres d'un auteur, si celui-ci a évolué dans sa façon d'écrire, donc dans sa façon de penser. On peut même aller jusqu'à établir la «paternité» (ou le manque d'indices de paternité) d'un texte. Cela est particulièrement intéressant dans le cas d'œuvres anciennes attribuées à certains auteurs ou pour les passages douteux. A ce sujet, cf. S. Allén et J. Thavenius (1970).

10. Restrictions

Comme n'importe quel instrument de travail, les concordances ont aussi leurs limites. On peut avancer que plus les concordances sont générales, plus elles servent de domaine de recherches et plus elles exigent, de la part du chercheur, un travail précis de remaniement des données. D'un autre côté, plus elles sont spécifiques, plus elles ont de restrictions dans leurs emplois. Il s'avère que la plupart des compilateurs de concordances ont un but bien défini avant de commencer leur travail de dépouillement. Si le but recherché est de nature littéraire, il sera plus ou moins bien adapté à des recherches d'une autre nature, mais ce n'est heureusement pas toujours le cas.

Appendice

1. Concordances et index intéressant les études romanes

Baudelaire: *A Concordance to Baudelaire's* Les Fleurs du Mal, edited by Robert T. Cargo. The University of North Carolina Press, Chapel Hill 1965.

Packard, D.: *A Concordance to Livy*. Harvard University Press, 1968.

Spevack, M.: *The Harvard Concordance to Shakespeare*. Cambridge, Mass. 1973.

Tombeur, P.: *Raoul de Saint-Trond. Gesta Abbatum Trudonensium* Livre IX. Index verborum, relevés statistiques. Hildesheim 1969.

Suzanne Hanon
Odense

Résumé

Les dépouillements comme les concordances et les index de mots sont présentés surtout sous leur aspect actuel d'inventaires produits de façon automatique sur ordinateurs. Une série de problèmes d'ordre méthodologique se posant au compilateur sont analysés: définitions du texte, du mot, concept de contexte, lemmatisation, indexation, etc. L'auteur montre, en s'appuyant sur les travaux les plus récents parus dans le domaine roman, l'emploi que l'on peut faire de ces inventaires de mots.

Bibliographie

Allen, J. R. (ed.): *The Study of French Literature with Computers*. University of Manitoba, Winnipeg 1973.

Allén, S. & J. Thavenius (red.): *Språklig databehandling. Datamaskinen i språk- og litteraturforskning*. Lund 1970.

Boone, A., H. Cuypers-Lippens, G. de Poerck, R. van Deyck-Bauwens, D. Willems, R. Zwaenepoel-Dhanis: «Projets et réalisations en traitement automatique dans le domaine du français». *Mélanges Imbs*, Strasbourg 1973; p. 329–341.

Borko, H. (ed.): *Automated Language Processing*. New York 1967.

Brackenier, R.: «Index et concordances d'auteurs français modernes. Etudes critiques». *Travaux de Linguistique* 3 (1972) 1–43 et 4 (1975) 1–61.

Burton, D. M.: «Some Uses of a Grammatical Concordance». *Computers and the Humanities* 2 (1968), 145–154.

Busa, R.: «Les travaux du 'Centro per l'Automazione dell'Analisi Letteraria' de Gallarate (Italie)», *Cahiers de Lexicologie* 3 (1962), 64–70.

Crosland, A. T.: «The Concordance and the Study of the Novel». *ALLC Bulletin* 3 (3) (1975) 190–196.

Cummings, L. A.: «The Electronic Humanist: Computing at Waterloo in Canada». *ALLC Bulletin* 3 (3) (1975) 226–234.

De Kock, J.: «De automatisering in de romaanse taalkunde». *Revue des Langues vivantes* 35 (1969) 175–193.

Des tracts en mai 68. Mesures de vocabulaire et de contenu (Demonet, M. *et al.*). Travaux et recherches de science politique n° 31, Paris 1975.

Dixon, J. E. G.: «A Prose concordance: Rabelais». *ALLC Bulletin* 2 (3) (1974) 47–54.

Dubois, J. *et al.*: *Introduction à la lexicographie: le dictionnaire*. Paris 1971.

Dubois, J. *et al.*: *Dictionnaire de linguistique*. Paris 1973.

Duggan, J. J.: «The Value of Computer-generated Concordances in Linguistic and Literary Research». *Revue du LASLA* no. 4 (1965) 51–60.

Packard, D.: *A Concordance to Livy*. Harvard University Press, 1968.

Spevack, M.: *The Harvard Concordance to Shakespeare*. Cambridge, Mass. 1973.

Tombeur, P.: *Raoul de Saint-Trond. Gesta Abbatum Trudonensium* Livre IX. Index verborum, relevés statistiques. Hildesheim 1969.

Suzanne Hanon
Odense

Résumé

Les dépouillements comme les concordances et les index de mots sont présentés surtout sous leur aspect actuel d'inventaires produits de façon automatique sur ordinateurs. Une série de problèmes d'ordre méthodologique se posant au compilateur sont analysés: définitions du texte, du mot, concept de contexte, lemmatisation, indexation, etc. L'auteur montre, en s'appuyant sur les travaux les plus récents parus dans le domaine roman, l'emploi que l'on peut faire de ces inventaires de mots.

Bibliographie

Allen, J. R. (ed.): *The Study of French Literature with Computers*. University of Manitoba, Winnipeg 1973.

Allén, S. & J. Thavenius (red.): *Språklig databehandling. Datamaskinen i språk- og litteraturforskning*. Lund 1970.

Boone, A., H. Cuypers-Lippens, G. de Poerck, R. van Deyck-Bauwens, D. Willems, R. Zwaenepoel-Dhanis: «Projets et réalisations en traitement automatique dans le domaine du français». *Mélanges Imbs*, Strasbourg 1973; p. 329–341.

Borko, H. (ed.): *Automated Language Processing*. New York 1967.

Brackenier, R.: «Index et concordances d'auteurs français modernes. Etudes critiques». *Travaux de Linguistique* 3 (1972) 1–43 et 4 (1975) 1–61.

Burton, D. M.: «Some Uses of a Grammatical Concordance». *Computers and the Humanities* 2 (1968), 145–154.

Busa, R.: «Les travaux du 'Centro per l'Automazione dell'Analisi Letteraria' de Gallarate (Italie)», *Cahiers de Lexicologie* 3 (1962), 64–70.

Crosland, A. T.: «The Concordance and the Study of the Novel». *ALLC Bulletin* 3 (3) (1975) 190–196.

Cummings, L. A.: «The Electronic Humanist: Computing at Waterloo in Canada». *ALLC Bulletin* 3 (3) (1975) 226–234.

De Kock, J.: «De automatisering in de romaanse taalkunde». *Revue des Langues vivantes* 35 (1969) 175–193.

Des tracts en mai 68. Mesures de vocabulaire et de contenu (Demonet, M. *et al.*). Travaux et recherches de science politique n° 31, Paris 1975.

Dixon, J. E. G.: «A Prose concordance: Rabelais». *ALLC Bulletin* 2 (3) (1974) 47–54.

Dubois, J. *et al.*: *Introduction à la lexicographie: le dictionnaire*. Paris 1971.

Dubois, J. *et al.*: *Dictionnaire de linguistique*. Paris 1973.

Duggan, J. J.: «The Value of Computer-generated Concordances in Linguistic and Literary Research». *Revue du LASLA* no. 4 (1965) 51–60.

Duggan, J. J.: «Formulas in the *Couronnement de Louis*». *Romania* 87 (1966) 315–344.
Duggan, J. J.: *The* Song of Roland: *Formulaic Style and Poetic Craft*. Berkeley 1973.
Engwall, G.: *A Concordance Program for Linguistic and Literary Research*. Computer Center Report no. 8, Royal College of Forestry, Stockholm 1972.
Engwall, G.: *Fréquence et distribution du vocabulaire dans un choix de romans français*. Stockholm 1974.
Fortier, P. A.: «Etat présent de l'utilisation des ordinateurs pour l'étude de la littérature française». *Computers and the Humanities* 5 (1971) 143–154.
Fossier L. et G. P. Zarri: *L'Indexation automatique des sources documentaires anciennes*. Éd. du CNRS, Paris 1975.
Frautschi, R. L.: «Recent Quantitative Research in French Studies». *Computers and the Humanities*, 7 (1973) 361–372.
Frautschi, R. L.: «An Overview of Recent Quantitative Research in French Studies», *in* Allen (ed.) 1973.
Geffroy, A., P. Lafon et M. Tournier: «Lexicometrical Analysis of Cooccurrences» *in* A. J. Aitken, R. W. Bailey and N. Hamilton-Smith: *The Computer and Literary Studies*. Edinburgh University Press, 1973, 113–134.
Glickman, R. J. and G. J. Staalman: *Manual for the Printing of Literary Texts and Concordances by Computer*. University of Toronto Press, 1966.
Gorcy, G., R. Martin, J. Maucourt, R. Vienney: «Le traitement des groupes binaires». *Cahiers de Lexicologie* 17 (1970) 15–46.
Grimal, P.: «Index» et «Concordances». *Revue des études latines* XLIV (1966) 108–116.
Hanon, S.: compte rendu de J. J. Duggan: *A Concordance of the Chanson de Roland* (Ohio State Un., 1969) et G. de Poerck *et al.*: *Le Charroi de Nîmes I–II* (Saint-Aquilin-de-Pacy, 1970). *Revue Romane* VIII, (1973) 421–423.
Hanon, S.: Chronique. «Colloque sur l'analyse des corpus linguistiques. Problèmes et méthodes de l'indexation maximale». *Cahiers de Lexicologie*, 23 (1973) 117–124.
Harris, Z. S.: «From Morpheme to Utterance». *Language* 22 (1946) 161–183.
Honeycutt, B. L.: «An Example of Comic Cliché in the Old French Fabliaux». *Romania* 96 (1975) 245–255.
Howard-Hill, T. H.: «On Literary Concordances, an early view». *ALLC Bulletin* 4 (3) (1976) 215–220.
IBM Data Processing Application «Literary Data Processing». New York 1971.
Ingram, W.: «Concordances in the Seventies». *Computers and the Humanities*, 8 (1974) 273–277.
Juilland, A.: *Dictionnaire inverse de la langue française*. Mouton, London – The Hague 1965.
Kersten, R.: «Las Concordancias de la *Obra Poética* de Eugenio Florit». Review. *Computer Studies in the Humanities and the Verbal Behavior* I (1968) 105–106.
Lamb, S. M. and L. Gould: *Concordances from Computers*. Mechanolinguistics Project, University of California, Berkeley, Cal. 1964.
Launay, M.: «Vocabulaire politique et vocabulaire religieux dans *Les Rêveries*». *Cahiers de Lexicologie* 5 (1964), 85–100.
Laurette, P.: «Concordances syntagmatiques et analyse de surface». *Computers and the Humanities* 8 (1974) 147–151.
Martinet, A.: «Le mot», *in*: *Problèmes de langage*, *Diogène* no. 51. Paris 1966, 39–53.
Michéa, R.: «De la relation entre le nombre des mots d'une fréquence déterminée et

celui des mots différents employés dans le texte». *Cahiers de Lexicologie* XVIII (1971) 65–78.

Michon, J. P. et M. Potdevin: «Recherche d'associations paradigmatiques et théorie des graphes». *Français moderne* 4 (1973), 433–447.

Mitterand, H. et J. Petit: «Index et Concordances dans l'étude des textes littéraires». *Cahiers de Lexicologie* 3 (1972), 160–175.

Moreau, R. (ed.): *Quelques applications de l'informatique en linguistique.* Etudes du développement scientifique, IBM France, Paris 1972.

Muller, Ch.: «Le mot, unité de texte et unité de lexique en statistique lexicologique». *Travaux de linguistique et de litt.* I (1963) 155–173.

Muller, Ch.: *Étude de statistique lexicale. Le vocabulaire du théâtre de Corneille.* Paris 1967.

Muller, Ch.: *Initiation à la statistique linguistique*, Paris 1968.

Muller, Ch.: «La lemmatisation. Essai d'analyse mathématique». *Travaux de linguistique et de litt.* XII (1974) 189–208.

Olsen, B. Munk: *Anvendelsen af elektronisk databehandling ved løsningen af filologiske opgaver: konkordanser, indices verborum.* Københavns Universitet, 1968.

Parrish, S. M.: «Concordance-making by Computer: It's Past, Future, Techniques, and Applications», *in Proceedings: Computer Applications to Problems in the Humanities.* A conversation in the disciplines, ed. by Frederick M. Burelbach, Brockport, N.Y. 1970, pp. 16–33.

Pottier, B.: *Introduction à l'étude des structures grammaticales fondamentales.* Publ. de la Faculté des Lettres, Nancy 1966.

Quemada, B.: *Les dictionnaires du français moderne 1539–1863.* Paris 1968.

Rasmussen, J.: «Facteurs déterminants de la combinaison sémantique d'éléments lexicaux». *Actes du 4e Congr. des Romanistes Scand. dédiés à H. Sten, Revue Romane* no. spec. 1 (1967) 129–138.

Rey, A.: *La lexicologie. Lectures.* Paris 1970.

Rey-Debove, J. (éd.): *La Lexicographie.* Langages 19 (1970).

Rey-Debove, J.: *Etude linguistique et sémiotique des dictionnaires français contemporains.* La Haye–Paris 1971.

Sáez-Godoy, L.: «Situation and Prospects of Computer-Aided Literary Research in Spanish». *Computers and the Humanities* 9 (1975) 245–246.

Sedelow, S. Y. and W. A. Sedelow: «Stylistic Analysis» *in* Borko (1967).

Spevack, M.: «Concordances: Old and New». *Computer Studies in the Humanities and Verbal Behavior* 4 (1973) 17–19.

Tasman, P.: «Indexing the Dead Sea Scrolls by Electronic Data Processing Methods». *IBM World Trade Corporation pamphlet*, New York 1958.

Tiefenbrun, S. W.: «Computational Stylistics and the French Classical Novel», *in*: Allen (ed.) (1973).

Togeby, K.: «Qu'est-ce qu'un mot?». *Recherches Structurales, TCLC* 5 (1949) 99–111.

Tombeur, P.: «Un faux problème: index ou concordance?». *Revue du LASLA* no. 2 (1967) 15–34.

Wagner, R. L.: *Les Vocabulaires français* I–II. Paris 1967–1970.

Zgusta, L.: *A Manual of Lexicography.* The Hague 1971.

raire se basent sur la même idée (Duggan, pour le style en formule des chansons de geste, *Romania*, 1966; Honeycutt, sur les clichés dans les fabliaux, *Romania*, 1975, etc.).

Il est aussi possible d'opérer n'importe quel *relevé statistique* à partir d'une concordance des groupes de deux, trois ou plusieurs autres éléments cooccurrents, mais cet aspect a déjà été mentionné plus haut.

9.3. Répartition

Si la concordance est bien préparée, elle permet l'étude de l'homogénéité du texte, donnant ainsi accès à la *répartition* des mots à travers le texte, l'œuvre, etc. Muller (1967) a bien montré pour l'œuvre de Corneille la répartition et l'évolution du vocabulaire chez cet auteur. On peut se demander, à partir d'une concordance à entrées alphabétiques ordonnées chronologiquement suivant les années de parution des différentes œuvres d'un auteur, si celui-ci a évolué dans sa façon d'écrire, donc dans sa façon de penser. On peut même aller jusqu'à établir la «paternité» (ou le manque d'indices de paternité) d'un texte. Cela est particulièrement intéressant dans le cas d'œuvres anciennes attribuées à certains auteurs ou pour les passages douteux. A ce sujet, cf. S. Allén et J. Thavenius (1970).

10. Restrictions

Comme n'importe quel instrument de travail, les concordances ont aussi leurs limites. On peut avancer que plus les concordances sont générales, plus elles servent de domaine de recherches et plus elles exigent, de la part du chercheur, un travail précis de remaniement des données. D'un autre côté, plus elles sont spécifiques, plus elles ont de restrictions dans leurs emplois. Il s'avère que la plupart des compilateurs de concordances ont un but bien défini avant de commencer leur travail de dépouillement. Si le but recherché est de nature littéraire, il sera plus ou moins bien adapté à des recherches d'une autre nature, mais ce n'est heureusement pas toujours le cas.

Appendice

1. Concordances et index intéressant les études romanes

Baudelaire: *A Concordance to Baudelaire's* Les Fleurs du Mal, edited by Robert T. Cargo. The University of North Carolina Press, Chapel Hill 1965.

Les Fleurs du Mal, Concordances, Index et relevés statistiques établis d'après l'édition Crépet-Blin par le Centre d'Etude du Vocabulaire Français de la Faculté des Lettres de Besançon avec la collaboration de K. Menemencioglu. *Documents pour l'étude de la langue littéraire*, publiés sous la direction de B. Quemada, Larousse, Paris 1965.

Concordance to Baudelaire's Petits Poèmes en Prose, with complete text of the poems compiled, and with an introduction by Robert T. Cargo. The University of Alabama Press, Alabama University 1971.

Bécquer: *A Concordance to the Poetry of Gustavo Adolfo Bécquer*, compiled, and with an introduction, by Enrique Ruiz-Fornells. The University of Alabama Press, Alabama University 1970.

Belli, G. G.: Federico Albano Leoni: *Concordanze belliane*, con lista alfabetica, lista inversa e rimario, I–III. *Acta Universitatis Gothenburgensia*, ed. H. Nilsson-Ehle, 1970.

Blondel de Nesle: *Les Chansons de Blondel de Nesle*. Concordances et index établis d'après l'édition L. Wiese par G. Lavis. Traitement automatique: C. Dubois. *Publications de l'Institut de Lexicologie française de l'Université de Liège*. Faculté de Philosophie et Lettres de l'Université de Liège, 1971.

Boccaccio: *Concordanze del* Decameron, a cura di Alfredo Barbina sotto la direzione di Umberto Bosco I–II. Accademia della Crusca, Firenze, C/E Giunti, G. Barbéra, 1969.

Chanson de Roland: *A Concordance of the* Chanson de Roland, compiled by Joseph J. Duggan. Ohio State University Press, 1969.

Charroi de Nîmes: Le Charroi de Nîmes, *chanson de geste*, par G. de Poerck, concordances, R. van Deyck, texte et variantes, R. Zwaenepoel, traitement automatique. Tome I–II. *Textes et traitement automatique*. Librairie-Editions Mallier, Saint-Aquilin-de-Pacy (Eure) 1970.

Chrétien de Troyes: *Philomena*. Concordances et index établis d'après l'édition C. de Boer, par C. Dubois, M. Dubois-Stasse et G. Lavis. Faculté de Philosophie et Lettres de l'Université de Liège. *Publications de l'Institut de Lexicologie française de l'Université de Liège*, 1970.

Guillaume d'Angleterre. Concordances et index établis d'après l'édition M. Wilmotte par M. Dubois-Stasse, A. Fontaine-Lauve. Traitement automatique: C. Dubois, M. Graitson, I–II. Faculté de Philosophie et Lettres de l'Université de Liège. *Publications de l'Institut de Lexicologie française de l'Université de Liège*, 1970.

Dante Alighieri: *A Concordance to the* Divine Comedy *of Dante Alighieri*. Edited by the Dante Society of America by Ernest Hatch Wilkins and Tho-

mas Goddard Bergin, Associate editor: Anthony J. de Vito. The Belknap Press of Harvard University Press, Cambridge, Mass. 1965.

Concordanza della Commedia *di Dante Alighieri.* Lovera, L. ed. I–III. Torino, Einaudi, 1975.

Joachim du Bellay: Joachim du Bellay. *La Deffence et Illustration de la Langue Francoyse.* Concordance établie par Suzanne Hanon. Traitement automatique: Poul Bonne Jørgensen et Ulf Hagen Køllgaard. *Etudes romanes de l'Université d'Odense*, vol. 6, Odense University Press, 1974.

García Lorca, Federico: *A Concordance to the Plays and Poems of Federico García Lorca.* Ed. Alice M. Pollin. Ithaca and London: Cornell University Press, 1975.

Garcilaso de la Vega: *Concordancias de las* obras poéticas *en castellano de Garcilaso de la Vega*, recopiladas por Edward Sarmiento en la edición de Elias L. Rivers. Editorial Castalia, Madrid 1970.

Florit, Eugenio: *Concordancias de la* Obra Poética *de Eugenio Florit.* Alice M. Pollin editora y compiladora. Parte I: *Concordancias.* Parte II: *Obra Poética.* New York University Press, New York, University of London Press, Limited, 1967. (Realizada en Institute for Computer Research in the Humanities, New York University, codificación electrónica por Henry Weitzer), *New York University Concordance Series, Spanish Literature.*

La Fontaine: *A Concordance to the* Fables *and* Tales *of Jean de La Fontaine*, edited by J. Allen Tyler. Cornell University Press, Ithaca and London 1974.

Leopardi: *Concordanze dei* Canti *del Leopardi.* Bufano, A., ed., Firenze 1969.

Lope de Rueda: Sáez-Godoy, Leopoldo: *El léxico de Lope de Rueda.* Clasificaciones conceptual y estadística. Bonn 1968.

Manzoni: *Concordanze degli* Inni sacri *di A. Manzoni.* A cura dell' Accademia della Crusca. Firenze 1967.

Pascal: *A Concordance to the* Pensées *of Pascal*, edited by Hugh M. Davidson and Pierre H. Dube. (Lafuma edition of 1952). Cornell University Press, Ithaca and London 1976.

Petrarca: *Concordanza delle* Rime *di Francesco Petrarca*, McKenzie, K. ed., Oxford 1912 (reprint: Torino, Bottega d'Erasmo, 1962).

Concordanze del Canzoniere *di Francesco Petrarca*, a cura dell'Ufficio lessicografico, vol. 1 + 2. Accademia della Crusca, *Opera del vocabolario*, Firenze 1971.

Porta, Carlo, Concordanze delle Poesie milanesi *di Carlo Porta*, a cura di S. Cipriani, Milano-Napoli 1970.

Racine: Bryant C. Freeman and Alan Batson: *Concordance du* théâtre *et des* poésies *de Jean Racine.* Cornell University Press, Ithaca, N.Y., vol. I–II, 1968.

Ronsard, Pierre de: *A Word-Index to the* Poetic Works *of Ronsard*, by A. E. Creore, I–II. *Compendia Computer-Generated Aids to Literary and Linguistic Research*, vol. 5, part I and II. Leeds 1972.

Rousseau: Index du Contrat Social (texte de 1762 et Manuscrit de Genève) par Michel Launay et Gunnar von Proschwitz, Collection des index et concordances des œuvres de Jean-Jacques Rousseau, série B, index des Œuvres, vol. I., Genève-Paris, 1977.

Textes italiens divers: Spogli elettronici dell'italiano delle origini e del duecento (SEIOD) et Spogli elettronici dell' italiano letterario contemporaneo (SEILC), en cours de publication, Societá editrice il Mulino, Bologna.

Verga: Marchi, G. P.: *Concordanze verghiane.* Verona 1972.

Villon: *François Villon*: *Œuvres d'après le manuscrit Coislin.* Rika van Deyck: Texte, variantes et concordances. Romana Zwaenepoel: Traitement automatique. I–II, Mallier, Saint-Aquilin-de-Pacy 1974.

Villehardouin: Villehardouin: *La Conqueste de Constantinople.* Cahiers du CRAL:

numéro 19: *Relevé des formes du manuscrit O de la Conqueste de Constantinople*, Nancy 1972.

numéro 20: *Index complet du manuscrit O de la Conqueste de Constantinople*, Nancy 1972.

numéro 22: *Index général des groupes nominaux dans le manuscrit O de la Conqueste de Constantinople*, Nancy 1973.

2. Autres travaux

Bompois, Cl.: *Concordance des Quatre Evangiles.* Tours 1965.

Busa, R. S. J. et A. Zampolli: *Concordantiae Senecanae* I–II. Hildesheim, New York 1975.

Ellison, J. W. (ed.): *Nelson's Complete Concordance of the Revised Standard Version of the Bible.* New York 1957.

Fernández Gomez, C.: *Vocabulario de Cervantes.* Real Academia Española, Madrid 1962.

Fernández Gomez, C.: *Vocabulario completo de Lope de Vega* I–III. Real Academia Española, Madrid 1971.

Grindstead, E.: *English Index to the Chinese Classics.* Lund 1975.

Hamesse, J.: *Thesaurus Bonaventurianus* I–IV. Louvain, Presses du CETEDOC, 1972–1975.

Holmboe, H.: *Concordance to Aeschylus'* Prometheus Vinctus. Akademisk Boghandel, Aarhus 1971.

McKinnon, A.: *The Kierkegaard Indices to* Kierkegaards Samlede Værker I–IV. Brill, Amsterdam 1970–1975.

Packard, D.: *A Concordance to Livy*. Harvard University Press, 1968.

Spevack, M.: *The Harvard Concordance to Shakespeare*. Cambridge, Mass. 1973.

Tombeur, P.: *Raoul de Saint-Trond. Gesta Abbatum Trudonensium* Livre IX. Index verborum, relevés statistiques. Hildesheim 1969.

Suzanne Hanon
Odense

Résumé

Les dépouillements comme les concordances et les index de mots sont présentés surtout sous leur aspect actuel d'inventaires produits de façon automatique sur ordinateurs. Une série de problèmes d'ordre méthodologique se posant au compilateur sont analysés: définitions du texte, du mot, concept de contexte, lemmatisation, indexation, etc. L'auteur montre, en s'appuyant sur les travaux les plus récents parus dans le domaine roman, l'emploi que l'on peut faire de ces inventaires de mots.

Bibliographie

Allen, J. R. (ed.): *The Study of French Literature with Computers*. University of Manitoba, Winnipeg 1973.

Allén, S. & J. Thavenius (red.): *Språklig databehandling. Datamaskinen i språk- og litteraturforskning*. Lund 1970.

Boone, A., H. Cuypers-Lippens, G. de Poerck, R. van Deyck-Bauwens, D. Willems, R. Zwaenepoel-Dhanis: «Projets et réalisations en traitement automatique dans le domaine du français». *Mélanges Imbs*, Strasbourg 1973; p. 329–341.

Borko, H. (ed.): *Automated Language Processing*. New York 1967.

Brackenier, R.: «Index et concordances d'auteurs français modernes. Etudes critiques». *Travaux de Linguistique* 3 (1972) 1–43 et 4 (1975) 1–61.

Burton, D. M.: «Some Uses of a Grammatical Concordance». *Computers and the Humanities* 2 (1968), 145–154.

Busa, R.: «Les travaux du 'Centro per l'Automazione dell'Analisi Letteraria' de Gallarate (Italie)», *Cahiers de Lexicologie* 3 (1962), 64–70.

Crosland, A. T.: «The Concordance and the Study of the Novel». *ALLC Bulletin* 3 (3) (1975) 190–196.

Cummings, L. A.: «The Electronic Humanist: Computing at Waterloo in Canada». *ALLC Bulletin* 3 (3) (1975) 226–234.

De Kock, J.: «De automatisering in de romaanse taalkunde». *Revue des Langues vivantes* 35 (1969) 175–193.

Des tracts en mai 68. Mesures de vocabulaire et de contenu (Demonet, M. *et al.*). Travaux et recherches de science politique n° 31, Paris 1975.

Dixon, J. E. G.: «A Prose concordance: Rabelais». *ALLC Bulletin* 2 (3) (1974) 47–54.

Dubois, J. *et al.*: *Introduction à la lexicographie: le dictionnaire*. Paris 1971.

Dubois, J. *et al.*: *Dictionnaire de linguistique*. Paris 1973.

Duggan, J. J.: «The Value of Computer-generated Concordances in Linguistic and Literary Research». *Revue du LASLA* no. 4 (1965) 51–60.