UNIVERSITE DE PROVENCE
ANNEE UNIVERSITAIRE 1999-2000

DIPLOME D'ETUDES APPROFONDIES
LANGAGE ET PAROLE
MENTION TRAITEMENT AUTOMATIQUE DES LANGUES

CHRISTOPHE REY

BIBLIOGRAPHIE ANALYTIQUE :
(Sous la direction de André VALLI)

« Dictionnaires électroniques : dictionnaires informatisés ou dictionnaires-machines ? »

Que nous l'envisagions comme un outil pédagogique, comme un outil culturel ou tout simplement comme une oeuvre littéraire, c'est-à-dire sous n'importe laquelle de ses multiples facettes, le dictionnaire, dont l'acception fait d'ailleurs l'objet d'un brillant article d'Etienne Brunet intitulé « Le mot *Dictionnaire* »[1], apparaît indissociable de la notion de représentation de la langue.

C'est en effet du besoin de représenter la langue, et les divers états de cet objet sans cesse en évolution, que semble être né le mouvement d'élaboration des dictionnaires et donc la longue tradition lexicographique française illustrée par des ouvrages comme le *Thresor de la langue françoyse* de Jean Nicot, le *Dictionnaire de l'Académie*, ou, plus près de nous, les dictionnaires *Larousse* et *Robert*.

Ce type d'ouvrages, dont Jean et Claude Dubois nous donnent une édifiante présentation dans leur « Introduction à la lexicographie : le dictionnaire »[2], a indubitablement connu au fil des siècles de nombreuses évolutions, parmi lesquelles figure une révolution sans précédent qui s'incarne à travers la métamorphose du livre, du Volumen, c'est-à-dire le rouleau, à celui de Codex, ouvrage rectangulaire.

Depuis plusieurs années, le dictionnaire est en passe de connaître une métamorphose tout aussi importante, grâce au développement de l'outil informatique, qui en plus d'offrir de nouveaux supports à ce dernier par le biais des CD-ROM ou de l'Internet, offre un nouveau visage à la lexicographie.

Nous allons au cours de ce travail nous attacher à présenter l'émergence de l'informatisation des dictionnaires en nous attardant sur la description des divers mouvements de ce processus. Nous ferons donc une distinction entre les dictionnaires électroniques réalisés à partir de dictionnaires papiers déjà existant, en évoquant au préalable la notion de « norme de codage » inhérente au phénomène de rétroconversion, les dictionnaires électroniques élaborés à partir de grands corpus (informatisés ou non), et les dictionnaires électroniques au sens profond du terme, c'est-à-dire les dictionnaires-machines uniquement destinés à être exploités par l'ordinateur.

---

[1] *E.L.A, Revue de Dialectologie des langues cultures. Hommage à Bernard Quemada « Dictionnairique et dictionnaires », 85-86, janv-juin 1992, Didier Erudition,*

[2] Ouvrage publié chez Larousse dans la revue *Langue et Langage* en 1971.

Le premier type de dictionnaires électroniques auxquels nous nous intéressons, en ne faisant aucune distinction entre les dictionnaires dits « de langue » et les dictionnaires « encyclopédiques », ou entre les dictionnaires bilingues et monolingues, est celle des dictionnaires informatisés.

Dans cette catégorie figurent uniquement des volumes ayant subi une « rétroconversion » : ils existaient déjà sous une forme papier et ont subi un traitement pour être disponibles sur support informatique. Dans son article « Rendre les dictionnaires plus actifs »[3], Gaston Gross fournit une définition très claire de la notion de dictionnaires informatisés, et établit une nette différence entre ce type d'ouvrage et ceux que nous aborderons plus loin, les dictionnaires-machines.

D'un point de vue historique, l'émergence de ce courant, que nous pouvons situer au début des années 1980, semble répondre, à une époque où le développement de l'informatique et du « Personnal Computer » sont sans cesse croissants, à un certain nombre de besoins pratiques, parmi lesquels se trouve indéniablement la nécessité d'offrir aux utilisateurs de dictionnaires une plus grande simplicité et disponibilité de consultation des ouvrages, souvent difficilement accessibles, tout au moins en ce qui concerne les dictionnaires anciens, notamment en raison de leur nombre réduit et de leur ancienneté.

Etroitement lié à ce besoin d'une consultation facilitée des ouvrages par l'utilisateur, un autre besoin, plus propre au lexicographe, se traduit par une volonté de disposer de toutes les informations du texte par le biais des recherches offertes par l'outil informatique.

Ce dernier propose en effet de grandes disponibilités d'exploitation des documents, bien plus efficaces et plus rapides que celles que restitue le travail manuel du lexicographe, mais intrinsèquement dépendantes du respect d'un certain nombre de règles qui lui sont propres. Ainsi, l'informatisation des dictionnaires papiers nécessite la mise en place de protocoles indispensables, parmi lesquels figure un élément de toute première importance, à savoir le choix de normes d'encodage des documents.

Il existe aujourd'hui tout un panel de normes d'encodage des dictionnaires, mais certaines d'entre-elles se sont peu à peu imposées. C'est le cas par exemple de la norme SGML (Standard Generalized Markup Language), norme ISO 8879, dont nous trouvons une présentation dans l'ouvrage publié par l'International Organisation for Standardization, *Langage normalisé de balisage généralisé (SGML)*, ISO 8879-1986 (F), Genève, 1986, mais aussi dans l'article de Lou Burnard, « What is SGML and how does it help ? »[4], et dans celui de Michel Goossens, 1995, « Introduction pratique à SGML »[5], qui couvre une partie importante des dictionnaires à ce jour informatisés, et ce notamment grâce à son adaptation par la Text Encoding Initiative (TEI), organisme créé par Nancy Ide. Cette dernière fournit d'ailleurs une présentation de celui-ci dans son article co-publié avec Jean Véronis, 1996, « Présentation de la TEI : Text Encoding Initiative »[6], qui prône l'utilisation de « Définitions de Type de Documents » (DTD), de véritables grammaires de documents.

La communication de Nancy Ide et Jean Véronis, parue en 1995 et intitulée « Encodage des dictionnaires électroniques : problèmes et propositions de la TEI »[7], nous donne un aperçu de la DTD TEI SGML élaborée pour les dictionnaires.

La norme SGML a, à titre d'exemple, été adoptée pour l'élaboration du CD-ROM (CD-ROM PC, version 1.0) *« Les dictionnaires des XVI^e^-XVII^e^ siècles »*, paru en 1998 chez Champion électronique, et qui comporte les versions informatisées de nombreux ouvrages anciens tels que le *Dictionnaire François* (1680) de César-Pierre Richelet, le *Dictionnaire Universel* (1690) d'Antoine Furetière, le *Dictionnaire de l'Académie Françoyse* (1694), ou le *Thresor de la langue francoyse* (1606) de Jean Nicot.

---

[3] *Lexicographie et informatique*, Autour de l'informatisation du *Trésor de la Langue Française* (Actes du Colloque International de Nançy, 29, 30, 31 mai 1995).
[4] In Ide and Véronis (1995) (Eds.) *The Text Encoding Initiative : Background and Context*, Kluwer Academic Publishers, Dordrecht, 1995, pp. 41-50,
[5] In *Cahiers GUTenberg*, n° 19, janvier 1995, pp. 27-58.
[6] In *TEI : Text Encoding Initiative, Cahiers de GUTenberg* n° 24, p.5.
[7] In *Lexicographie et informatique*, Autour de l'informatisation du *Trésor de la Langue Française* (Actes du Colloque International de Nançy, 29, 30, 31 mai 1995),

Citons également l'Extensible Markup Language (XML), langage alliant la richesse syntaxique de la norme SGML, norme de laquelle il est d'ailleurs dérivé, et la simplicité de l'Hypertext Markup Language (HTML), langage de balisage des informations sur le World Wide Web, qui se trouve présenté dans l'ouvrage de Michard Alain, *XML Langage et applications*, publié en 1998 chez Eyrolles, ainsi que dans le volume de Tim Bray, Jean Paoli et C.M.Sperberg-McQueen, 1998, *Langage de balisage extensible (XML) 1.0 Recommandation du W3C, 10 février 1998*, et qui est indubitablement appelé à devenir l'une des normes de codage de l'avenir. Son adaptation par la Text Encoding Initiative, dont nous trouvons une présentation dans l'ouvrage *Construction of an XML version of the TEI DTD* de C.M. Sperberg-Mc Queen, disponible sur Internet à l'adresse suivante : http://www.uic.edu/orgs/tei/ed/edw69.html, et celui de Robin Cover, intitulé Text Encoding Initiative (TEI) – XML for TEI Lite, également consultable sur Internet, plus précisément à l'adresse http://www.oasis.open.org/cover/tei.html.

Face à ce type de balisage, certains projets d'informatisation privilégient le recours à des balisages « propriétaires », c'est-à-dire des langages de balisages n'étant pas des normes à part entière mais des produits développés par des organismes privés.

C'est le cas des logiciels Text Analysis Computing Tools (TACT), logiciel élaboré à l'Université de Toronto, et WORDCRUNCHER, mis au point à la Brigham Young University, dont nous trouvons une brève présentation dans l'article d'Agnès Tutin, Chantal Wionet et Nathalie Lanckriet, 1999, « SGML pour l'informatisation des dictionnaires anciens : l'expérience du Dictionnaire Universel de Furetière revu par Basnage (1702) »[8] respectivement exploités par le lexicographe T.R. Wooldridge pour l'encodage du *Trésor de la Langue Française* de Jean Nicot et l'équipe constituée de Philippe Caron, Louise Dagenais et Gérard Gonfroy pour la constitution d'une base minimale du *Dictionnaire Critique* (1787) de l'Abbé Féraud.

[8] Article paru dans la revue publiée par l'Institut de Recherche et d'Histoire des Textes, *Le médiéviste et l'ordinateur*, plus précisément dans le numéro 38[8], (Hiver 1999) intitulé « Le texte médiéval sur Internet (2). Mettre des textes sur Internet»,

Dans notre perspective d'illustration du mouvement d'informatisation des dictionnaires, l'évocation des travaux des lexicographes cités nous amène tout naturellement à faire une distinction entre les divers dictionnaires informatisés, et en particulier à établir une distinction entre les ouvrages anciens et les ouvrages modernes.

Le mouvement d'informatisation des dictionnaires, bien qu'ayant historiquement débuté avec la rétroconversion du *Thresor de la langue françoyse* (1606) de Jean Nicot, est plus particulièrement, et ce certainement pour des raisons pratiques, représenté par l'informatisation de répertoires modernes. Au cours des dix dernières années, ces derniers ont en effet été nombreux à travers le monde, à avoir subi une rétroconversion. Soulignons d'ailleurs que ce mouvement ne s'est pas cantonné aux seuls dictionnaires de langue, mais qu'il s'est étendu aux dictionnaires spécialisés, ainsi qu'en témoignent les nombreux ouvrages disponibles sur Internet, à partir de sites comme *Leximagne - l'empereur des pages dico@ Globe Gate,* dont l'adresse est la suivante: http://www.utm.edu/departments/french/dico.html

Le monopole de production des dictionnaires de langue étant en France essentiellement détenu par les trois grandes maisons d'édition que sont Larousse, Robert et Hachette, il n'est donc pas étonnant de constater que c'est sous l'égide de ces dernières que sont placés les principaux dictionnaires modernes de langue informatisés.

Avant de présenter quelques-uns de ces ouvrages modernes, il semble important de souligner qu'en ce qui concerne l'édition de versions informatisées sur CD-ROM de plusieurs dictionnaires anciens, elle n'est pas assurée par ces éditeurs mais par exemple par Champion, à qui l'on doit entre autres le *Dictionnaire de l'Académie française – édition de 1694*, CD-ROM PC/MAC, Champion électronique, 1998, *« Les dictionnaires des XVI^e^-XVII^e^ siècles »*, CD-ROM PC, version 1.0, Champion électronique, 1998, évoqué plus haut, ou les dictionnaires à paraître, *Grand dictionnaire universel du XIX^e^* de Pierre Larousse, CD-ROM PC/MAC, Champion électronique, 2000, et *Les Dictionnaires de l'Académie française (XVII^e^ et XVIII^e^ siècles)*, CD-ROM PC, Champion électronique, 2000.

Parmi les principaux ouvrages modernes informatisés, nous pouvons mentionner les versions électroniques de certains répertoires parus chez Robert, tels que *Le Robert électronique* publié en 1990, regroupant le texte intégral des neufs volumes du Grand Robert de la langue française nouvelle édition, ainsi qu'un dictionnaire de citations, *Le Grand Robert électronique*, Dictionnaires Le Robert, 1994, ou *Le petit Robert. Dictionnaire analogique et alphabétique de la langue française : Nouvelle éd. du Petit Robert de Paul Robert*, texte remanié et amplifié sous la direction de Josette Rey-Debove et Alain Rey, CR-ROM PC/MAC, version 1.2, Dictionnaires Le Robert, 1996. Notons qu'une description rapide des fonctionnalités du Robert électronique nous est faite par Etienne Brunet dans son article publié en 1997 et intitulé « Les dictionnaires électroniques »[9].

Citons également la publication en 1998, chez Larousse, du *Bibliorom Larousse* comportant 6 ouvrages : *Le petit Larousse illustré*, le *Thésaurus*, le *Dictionnaire des citations françaises et étrangères*, le *Dictionnaire compact*

[9] In *Revue française de linguistique appliquée, 1997, II-1* (7-30).

français-anglais/anglais-français, le *Dictionnaire compact français-allemand/allemand-français,* et le *Dictionnaire compact français-espagnol/espagnol-français,* ainsi que la parution chez Hachette du *Dictionnaire Hachette encyclopédique 2000,* 2 CD-ROM ou 1 DVD PC & MAC, Hachette Multimédia, 1999.

Ce dernier ouvrage nous amène à succinctement évoquer une nouvelle perspective dans l'informatisation des dictionnaires modernes, et qui consiste en l'élaboration de nouveaux répertoires essentiellement électroniques, autrement dit n'existant pas au préalable sous une forme papier. Cette entreprise, bien que divergeant donc du mouvement de rétroconversion des dictionnaires, n'est cependant pas assimilable au principe d'élaboration des dictionnaires-machines que nous décrirons plus loin, notamment en raison du fait qu'il s'agit toujours d'ouvrages mis à la disposition de lecteurs humains et non d'un ordinateur.

En ce qui concerne les publications de dictionnaires modernes parus dans des pays francophones, mentionnons les travaux de l'équipe pour le *Trésor de la Langue Française au Québec* (TLFQ) qui travaille à l'élaboration d'une version informatisée du *Dictionnaire du Français Québécois* (DFQ), ouvrage dont la conception a débuté dans les années 1970, et ce grâce aux logiciels WordPerfect, Ventura Publisher et TACT. Nous trouvons d'ailleurs une présentation de cette entreprise dans la communication de Alain Auger et Claude Poirier, 1994, « L'exploitation du *Dictionnaire du français québécois* au moyen du logiciel TACT »[10].

[10] 1994 ,CCH Working Papers 4.

La rétroconversion des dictionnaires anciens, est un phénomène qui trouve ses origines au début des années 1980, introduit par les travaux du lexicographe Terence Russon Wooldridge sur le *Thresor de la langue françoyse* de Jean Nicot, et qui depuis connaît un succès grandissant.

Dans une toute autre perspective que celle de la simple numérisation des textes anciens, dont nous trouvons un aperçu dans l'ouvrage *Les documents anciens*. Document numérique, Volume 3. n°1-2 juin 1999, Hermes Science Publication, Paris, qui répond essentiellement à un besoin de possession des ouvrages sur support informatique pour notamment assurer leur pérennité et leur disponibilité, le mouvement d'informatisation des dictionnaires anciens constitue certes une réponse à ce même besoin de conservation des documents, mais repose également sur un besoin de consultation plus aisée et plus performante des articles dictionnairiques et de leur contenu par le biais d'un étiquetage d'une partie ou de la totalité des informations qui les composent.

L'article « Ce que les linguistes peuvent attendre d'un dictionnaire informatisé »[11], rédigé par Danielle et Pierre Corbin, et Agnès Tutin (avec la participation de Sophie Aliquot), nous donne d'ailleurs un aperçu des disponibilités offertes à des spécialistes par la version informatisée d'un dictionnaire.

De nombreux ouvrages anciens, parmi lesquels figurent, pour ne citer que quelques-uns des plus célèbres, le *Dictionnaire de l'Académie*, le *Thresor de la langue françoyse* de Nicot, le *Dictionnaire Etymologique ou Origines de la Langue française* de Gilles Ménage, le *Dictionnaire Critique* de Jean-François Féraud (1787), et le *Dictionnaire Universel* revu par Basnage de Bauval (1702), ont ainsi connu une rétroconversion.

En ce qui concerne le dernier de ces ouvrages, sa rétroconversion n'est pas encore effective, mais en élaboration au sein du projet IDEA (Informatisation des Dictionnaires Encyclopédiques Anciens), qui prévoit également le passage sur

[11] Article paru dans le volume *Lexicographie et informatique*, Autour de l'informatisation du *Trésor de la Langue Française* (Actes du Colloque International de Nançy, 29, 30, 31 mai 1995)

support informatique du Dictionnaire Universel d'Antoine Furetière (1690) et de la première édition du Trévoux (1704), dirigée par Chantal Wionet du laboratoire Syntaxe Interprétation Lexique (LISL). Une présentation de ce projet se trouve d'ailleurs sur Internet à l'adresse suivante : http//www.univ-lille3.fr/www/silex/wionet/IDEA.site.html.

L'informatisation de l'édition du *Dictionnaire Universel* proposée par Basnage de Bauval fait quant à elle l'objet de l'ouvrage à paraître de Chantal Wionet et Agnès Tutin, *Informatisation du Dictionnaire Universel de Furetière revu par Basnage de Bauval (1702) : premier bilan*, Paris, Honoré Champion. Un échantillon de ce travail est notamment disponible dans les actes du Colloque DictA1998, *Table ronde sur l'informatisation des dictionnaires anciens, Limoges, 19-20 novembre 1998*, colloque organisé par le Groupe d'Étude en Histoire de la Langue Française (G.E.H.L.F) et la Société Internationale d'Etudes Historiques et Linguistiques des Dictionnaires Anciens (SIEHLDA).

L'informatisation de l'œuvre de Gilles Ménage est, elle, entreprise par Isabelle Leroy-Turcan et notamment présentée dans son article, 1994, « L'informatisation du Dictionnaire Etymologique ou Origines de la Langue Françoise (1694) de Gilles Ménage, 1694 : »[12], dans lequel elle présente à la fois le fameux dictionnaire ainsi que les aspects et les intérêts de son informatisation, et dans sa communication lors des *Actes des journées « Dictionnaires électroniques du français des XVIème et XVIIème siècles »*, organisées à Université Blaise Pascal, Clermont-Ferrand, les 14-15 juin 1996, intitulée « Intérêt d'une base informatisée pour le Dictionnaire Etymologique ou Origines de la Langue Françoise, 1694, de Gilles Ménage : les modalités de mise en oeuvre ».

Mentionnons également la remarquable présentation de l'ouvrage de Ménage à travers les Actes du colloque organisé par Isabelle Leroy-Turcan et T.R.Wooldridge : *« Gilles Ménage (1613-1692), grammairien et lexicographe. Le*

rayonnement de son oeuvre linguistique ». Actes du colloque international tenu à l'occasion du tricentenaire du *Dictionnaire étymologique ou Origines de la langue françoise* (1694), Université Jean Moulin Lyon III, 17-19 mars 1994.

Malgré le nombre restreint des travaux dont il a fait l'objet, l'ouvrage de Jean Nicot est, avec celui de l'Académie Française, l'un des dictionnaires anciens qui semble avoir le plus suscité l'attention des lexicographes. C'est à Térence Russon Wooldridge, de l'Université de Trinity Collège à Toronto, véritable artisan de l'informatisation des dictionnaires, que le *Thrésor de la langue françoyse* (1606) doit néanmoins sa version informatisée, comme en témoignent les diverses publications de ce dernier. Citons notamment la communication parue en 1998 dans les actes du Colloque DictA1998, *Table ronde sur l'informatisation des dictionnaires anciens, Limoges, 19-20 novembre 1998,* et intitulée *La rétroconversion d'Estienne et de Nicot pour mise sur Internet,* dans laquelle l'auteur évoque les différentes étapes de l'élaboration de la version informatisée du *Thrésor* et les modalités de cette entreprise.

Placée sous l'égide des lexicographes Isabelle Leroy-Turcan et T.R.Wooldridge, l'informatisation du *Dictionnaire de l'Académie Française* fait l'objet de nombreuses publications parmi lesquelles nous pouvons citer le colloque DictA1998, *Table ronde sur l'informatisation des dictionnaires anciens, Limoges, 19-20 novembre 1998*, organisé en collaboration entre le GEHLF et la SIEHLDA et auquel nous pouvons rattacher la communication d'Isabelle Leroy-Turcan et Russon Wooldridge: « L'informatisation du *Dictionnaire de l'Académie française* », qui évoque les principales caractéristiques de ce projet.

Nous trouvons dans les « Modalités de mise en oeuvre de l'informatisation de la première édition du Dictionnaire de l'Académie française (1694) » d'Isabelle Leroy-Turcan, publiées dans les *Actes des Journées « Dictionnaires électroniques des XVIe-XVIIe s. », Clermont-Ferrand, 14-15 juin 1996*, un aperçu historique du projet d'informatisation des différents volumes du célèbre dictionnaire et les premières perspectives de balisage du texte.

---

[12] *Actes du Congrès International sur l'Informatisation des Dictionnaires anciens (Toronto,*

Russon Wooldridge, dans son «Projet d'informatisation du *Dictionnaire de l'Académie* (1694-1935)»[13], fournit également une description concise du travail d'informatisation du dictionnaire de l'Académie.

Aux vues des diverses publications que nous venons de présenter, il semble indubitable que l'informatisation des dictionnaires anciens est un phénomène d'une ampleur conséquente. Ce dernier semble d'ailleurs d'autant plus considérable dans la mesure où une grande partie de ces travaux sont rendus disponibles sur l'outil Internet et notamment grâce à l'existence d'associations, au sein desquelles sont regroupés bon nombre de laboratoires et de chercheurs, pour la promotion et la publication des recherches effectuées dans le domaine de la lexicographie ancienne.

A titre d'exemples, nous pouvons citer, la Société Internationale d'Etudes Historiques et Linguistiques des Dictionnaires Anciens (SIEHLDA) qui siège à l'Université de Lyon III et à laquelle nous devons la publication de nombreux actes de colloques parmi lesquels figurent certains de ceux précédemment cités, ou le Groupe d'Étude en Histoire de la Langue Française (G.E.H.L.F) qui fait partie de l'Institut National de la Langue Française et est dirigé par Philippe Caron. A ces organismes sont par ailleurs associés certains sites Internet qui offrent « en ligne » les éditions électroniques d'Actes de colloques, ouvrages ou compte-rendus, et des « bases » dictionnairiques, c'est-à-dire un ensemble d'éditions informatisées de quelques ouvrages anciens. Les sites dictA (Dictionnaires Anciens), EDICTA (Early Dictionaries, Dictionnaires Anciens) et du projet American and French Research on the Treasury of the French Language de l'Université de Chicago (ARTFL), sont par exemple en mesure d'offrir plusieurs bases dictionnairiques : La base *Dictionnaires d'autrefois*, placée sous l'égide de l'ARTFL, dans laquelle sont regroupés les ouvrages d'Etienne (1552),

---

octobre 1993) : Early Dictionary Databases, CCH Working Papers, 4, 1994, pp.131-142,

[13] Article paru dans les *Actes du Colloque international Le Dictionnaire de l'Académie française et la lexicographie institutionnelle européenne*, Institut de France, novembre 1994; (ed. B. Quemada & J. Pruvost), Paris, Champion: 309-20.

de Nicot (1606), et de l'Académie (1694, 1798, 1835), la base Nicot-Académie-Féraud, la Base Echantillon avec hypertexte analytique du *Thresor de la langue françoyse* 1606 de Nicot, ou la Base du dictionnaire intégral de ce même ouvrage et l'*Early Modern English Dictionaries Database*, base dictionnairique regroupant divers répertoires anglais anciens, respectivement mis en place par l'ARTLF et Ian Lancashire. Notons également l'existence de la riche « Base Echantillon des Dictionnaires Français Anciens », réunissant le *Thresor de la langue françoyse* (1606), le *Dictionnaire françois* (1680) et le *Dictionnaire portatif* (1784, version augmentée par Wailly) de César-Pierre Richelet, le *Dictionnaire Universel* d'Antoine Furetière (1690), le *Dictionnaire de l'Académie* (1694-1935), le *Dictionnaire de Trévoux* (1771) et le *Dictionnaire critique* (1687) de Jean-François Féraud.

Dans sa communication, 1998, «Les dictionnaires anciens sur Internet: bases linguistiques, philologiques, culturelles»[14], Térence Russon Wooldridge présente à la fois les modalités d'informatisation et de mise en base des ouvrages anciens, puis dresse un aperçu des divers types d'analyses permises par ces bases dictionnairiques hypertextuelles.

Malgré cet aspect de cohésion qui semble caractériser le mouvement d'informatisation des dictionnaires anciens, notamment grâce à la mise en place des bases dictionnairiques faisant cohabiter les divers projets, ce dernier est toutefois partagé entre deux perspectives sur le type d'encodage à adopter.

Le balisage « minimal formel » prôné par T.R.Wooldridge pour la rétroconversion du *Thresor de la langue françoyse* et du *Dictionnaire de l'Académie Française*, ainsi que par Louise Dagenais, Philippe Caron et Gérard Gonfroy pour l'informatisation du *Dictionnaire Critique*, privilégie un encodage réduit de la microstructure de l'article et permet la consultation des divers champs informationnels par le truchement de « mots-clefs métalinguistiques », dont nous trouvons une présentation dans l'article de T.R.Wooldridge et Isabelle Leroy-

[14] Communication préparée pour le XVIe Congrès international de l'Association Guillaume Budé, Limoges, août 1998.

Turcan, 1996, «Les mots-clés métalinguistiques comme outil d'interrogation structurante des dictionnaires anciens (le cas du *Dictionnaire de l'Académie française* par comparaison avec le *Thresor* de Jean Nicot et le *Dictionnaire Etymologique ou Origines de la Langue Françoise* de Gilles Ménage), »[15].

Soulignons également que ce mode de balisage fonctionne avec les logiciels propriétaires TACT ou WordCruncher, évoqués plus haut.

Le recours à ces deux logiciels est également retenu pour le balisage « analytique fin » proposé par Isabelle Leroy-Turcan lors de ses travaux sur l'informatisation du *Dictionnaire Etymologique ou Origines de la Langue Françoise*. Le balisage analytique fin d'Isabelle Leroy-Turcan, bien qu'en apparence relativement proche du balisage minimal, diffère de ce dernier dans la mesure où la consultation des divers champs compositionnels de l'article ne repose non plus sur l'utilisation de « mots-clefs métalinguistiques » mais sur celle de « séquences-clefs métalinguistiques », notion qu'elle nous présente dans son « Intérêt d'une base informatisée pour le *Dictionnaire Etymologique ou Origines de la Langue Françoise*, 1694, de Gilles Ménage: les modalités de mise en œuvre»[16], paru en 1996, et qu'il prévoit l'encodage de tous les champs informationnels de l'article dictionnairique.

En ce sens, le balisage analytique prôné par Isabelle Leroy-Turcan se rapproche de celui adopté par Chantal Wionet et Agnès Tutin pour la rétroconversion du *Dictionnaire Universel* de Furetière revu par Basnage de Bauval (1702). Il en est toutefois divergent par le fait que le «balisage analytique formalisé » repose non pas sur un balisage propriétaire proposé par les logiciels TACT ou WordCruncher, mais sur le Standard Généralised Markup Language, une norme internationale d'encodage et d'échange de documents.

Pour une présentation de ces deux types de balisage qui scindent le mouvement d'informatisation des dictionnaires anciens, nous pouvons nous référer à l'article d'Isabelle Leroy-Turcan, 1998, *«Balisage formel ou balisage*

[15] In *Lexicomatique et dictionnairiques* (éd. A. Clas, P. Thoiron & H. Béjoint), 1996, Beyrouth: FMA / Montréal: AUPELF-UREF: 307-16.

fin pour les dictionnaires anciens informatisés : objectifs et implications méthodologiques»[17], ou à l'article de T.R.Wooldridge 1997, *« Baliser un texte c'est le penser: le cas du Dictionnaire de l'Académie Française»*[18], dans lequel le lexicographe, en avançant la complexité du balisage analytique fin, prône clairement l'utilisation du balisage minimal.

A ce type de dictionnaires, c'est-à-dire les dictionnaires existant sous une forme papier et ayant subi une informatisation, nous voudrions sommairement rattacher une autre variété d'ouvrages, moins nombreux, qui sont eux aussi des dictionnaires papiers informatisés, mais qui diffèrent des premiers par le mode d'élaboration de leur version manuelle.

En effet, celle-ci repose sur une linguistique dite « de corpus », c'est-à-dire sur une linguistique essentiellement basée sur la constitution de grands corpus, outils destinés à fournir au lexicographe une approche différente du lexique, moins guindée et artificielle que celle consistant à construire lui-même, à partir d'une grande part de sa propre subjectivité, les articles dictionnairiques. Par le biais de ces gigantesques corpus, ce dernier possède alors un outil lui permettant d'élaborer et de classer les entrées de son dictionnaire en tenant compte de la récurrence des diverses lexies apparaissant dans les textes qu'il a sélectionnés. Cette méthode a également l'énorme avantage de lui permettre d'illustrer ces diverses lexies par des exemples tirés de situation de production du langage réelles.

Parmi les ouvrages les plus célèbres figurant dans cette catégorie de dictionnaires, se trouvent le *Trésor de la Langue Française*, communément connu sous le nom de TLF, de l'Institut National de la Langue Française (INALF) et le *Collins COBUILD English Dictionnary* de John Sinclair.

Le *COBUILD English Dictionnary*, est un dictionnaire pour apprenants étrangers, élaboré dans les années 1980 à l'Université de Birmingham, sous la

[16] Article publié dans les Actes des Journées « Dictionnaires électroniques des XVIe-XVIIe s». Clermont-Ferrand, 14-15 juin 1996.
[17] Actes du colloque DictA1998, Limoges, novembre 1998.

direction de John Sinclair. Sa grande particularité est, comme nous venons de le mentionner, de reposer sur une nouvelle approche de l'élaboration de l'outil dictionnaire, reposant sur l'utilisation d'un corpus linguistique. Ce dernier, baptisé en 1991 de « Bank of English » a été constitué depuis 1980 à partir de corpus aussi bien oraux qu'écrits, issus de livres, de magasines, de journaux, de conversions radiophoniques ou télévisuelles : en résumé, de sources essentiellement contemporaines, et constituant en 1998 une base d'environ 330 millions de mots, traitée par informatique. Dans la préface de l'édition de 1987 du *Collins COBUILD English language dictionary* édité par London Glasgow : Collins, John Sinclair résume les particularités de cet ouvrage.

L'élaboration du *Trésor de la Langue Française* de l'INALF, est un projet créé à l'initiative du philologue et linguiste Paul Imbs, lors du congrès du 16 novembre 1957. Ainsi que l'illustre Eveline Martin dans son article, 1990, « Sources et ressources du TLF. De la boîte à fiche au disque compact »[19], la mise en place de cette vaste entreprise répond à un profond désir de renouvellement de la lexicographie française des XIXe et XXe siècles. Ce renouveau, à la fois dû à la structure et au contenu du répertoire, prend néanmoins toute son ampleur à travers le mode de conception de ce dernier, qui à l'image du *COBUILD*, repose également sur l'utilisation d'un corpus linguistique.

Ce dernier, dont une présentation est faite dans l'article de Gérard Gorcy, 1990, « Le Trésor de la Langue Française (TLF). Son originalité et les voies ouvertes pour son informatisation »[20], se distingue comme la base textuelle informatisée la plus importante au monde. Elle fait à présent partie intégrante d'un immense corpus de textes français du XVIe au XXe siècles, composés de plus 3500 oeuvres, à 80% littéraires et à 20% techniques, et connu sous le nom de base FRANTEXT. Christine Ducourtieux, dans son article paru en 1998 et intitulé

---

[18] Version révisée d'une communication faite à Paris, le 23 mai 1997, dans le cadre d'une journée d'études organisée par le Groupe d'Études en Histoire de la Langue Française,

[19] Autour d'un dictionnaire : le « Trésor de la Langue Française », in *Dictionnairique et lexicographie 1-1990*.

[20] Autour d'un dictionnaire: le « Trésor de la Langue Française », in *Dictionnairique et lexicographie 1-1990*.

« Quelques bases TEXTUELLES »[21], nous donne un aperçu de cette base en présentant le corpus qui la constitue et le logiciel qui permet son exploitation, avant de fournir en annexe la liste des ouvrages du XVIe siècle la composant.

Malgré la notoriété que lui confère son aspect d'œuvre de longue haleine réalisée sous l'égide du prestigieux INALF, le *Trésor de la Langue Française* n'a pas vraiment connu le succès escompté par ses créateurs. Ceci s'explique notamment par le fait qu'il constitue une oeuvre apparue en pleine période de « mutation » de la lexicographie, et donc occultée par le mouvement de généralisation des dictionnaires électroniques. Ceci apparaît comme d'autant plus vrai dans la mesure ou le TLF semble être voilé par sa propre version électronique : le *Trésor de la Langue Française Informatisé* (TLFI).

Egalement placé sous l'égide de l'INALF, le TLFI, ouvrage que nous aurions tout aussi bien pu évoquer lors de la présentation des répertoires modernes informatisés, est un projet colossal qui constitue une réponse au développement de la lexicographie informatique et bénéficie pour son élaboration, à la fois de l'existence de la base textuelle informatisée du TLF, et de l'expérience apportée par l'informatisation de *l'Oxford English Dictionary*. Sa rétroconversion a notamment fait l'objet d'un colloque international à Nancy, en 1995, dont les diverses communications ont été regroupées sous le volume *Lexicographie et informatique*, « Autour de l'informatisation du Trésor de la Langue Française », Actes du Colloque International de Nancy (29, 30, 31 mai 1995), Paris, Didier Erudition. Notons d'ailleurs que dans cet ouvrage, l'article, 1995, « Le projet d'informatisation du TLF » de Jacques Dendien, ainsi que son titre l'indique, dresse les caractéristiques du projet de rétroconversion du TLF. Plusieurs articles de Françoise Henry font également état de ces caractéristiques : c'est le cas de la communication, 1995, « Pour une informatisation du TLF », figurant dans l'œuvre *Lexicographie et informatique*, qui évoque les divers intérêts que présente l'informatisation de cet ouvrage, mais aussi des articles, 1990, « Informatisation

[21] Article paru dans la revue *Le médiéviste et l'ordinateur*, et plus particulièrement dans le numéro 37 de celle-ci (Hiver 1998), « Le texte médiéval sur Internet (1). Chercher et trouver ».

du Trésor de la langue française : problèmes et perspectives »[22] et, 1992, « Informatisation du TLF : problèmes, finalités, moyen (troisième approche) »[23].

Le troisième type de répertoires qui se distingue dans le mouvement d'informatisation des dictionnaires correspond à ce que nous pouvons nommer les dictionnaires électroniques, ou à proprement parler, les « dictionnaires-machines ». En d'autres termes les ouvrages n'existant que sur support électronique, n'ayant donc pas subi d'informatisation, et uniquement utilisables par des ordinateurs. Dans son article paru en 1998 sous le titre «Une autre lecture du dictionnaire de langue : le CD-ROM »[24], Christine Jacquet-Pfau dresse une présentation très claire de ce type d'ouvrages, en les distinguant notamment des dictionnaires ayant subi une rétroconversion. Notons que cette distinction avait déjà été ébauchée quelques années auparavant par Philippe Herr dans sa communication, 1991, « Les dictionnaires électroniques : quelles caractéristiques pour quels objectifs »[25].

L'émergence de ces ouvrages s'inscrit dans la mouvance du Traitement Automatique du Langage (TAL), discipline visant à traiter de façon « automatique », c'est-à-dire par le biais de l'outil informatique, les données linguistiques de tout genre. Le volume de Catherine Fuchs, *Linguistique et Traitement Automatique des Langues*, Hachette Université, 1993, fait office d'ouvrage de référence en ce qui concerne la présentation de cette discipline, au même titre d'ailleurs que les 2 tomes de l'œuvre de Gérard Sabah, *L'intelligence artificielle et le langage (Volume 1, Représentation des connaissances 2e édition)* et *L'intelligence artificielle et le langage (Volume 2, Processus de compréhension)*, respectivement parus en 1990 et 1989 aux éditions HERMES.

---

[22] Communication publiée dans le numéro *56-57*, 1990 1-2, des *Cahiers de lexicologie*, Didier Erudition.

[23] Communication publiée dans le numéro n° 61, 1992 – 2, des *Cahiers de lexicologie*, Didier Erudition.

[24] *La Tribune Internationale des Langues Vivantes*, n°24, Novembre 1998, pp.63-71.

[25] *La Tribune Internationale des Langues Vivantes*, n°7, Mai 1991, pp.19-1.

L'informatisation des dictionnaires papiers, bien que constituant une première approche dans le traitement du langage par ordinateur, étant donné qu'elle génère des outils uniquement consultables sur ordinateur, ne répond pas, ainsi que le déclarent Blandine Courtois et Max Silberztein dans leur article, 1990, « Dictionnaires électroniques du français »[26], à une véritable description formelle du langage. Aux dictionnaires informatisés, mis à la disposition de l'utilisateur humain doté d'une grande connaissance linguistique, s'opposent les dictionnaires électroniques, uniquement élaborés pour être utilisés par des programmes informatiques à priori dépourvus de toute connaissance linguistique. Face à l'incomplétude des dictionnaires usuels et à leur grande part d'implicite, se dresse la nécessaire complétude des dictionnaires électroniques susceptibles d'analyser tous les mots d'un texte et le caractère obligatoirement explicite de leurs informations. Soulignons également comme troisième caractéristique opposant les deux types d'ouvrages, l'aspect de cohérence structurelle des entrées qu'impose l'association étroite des dictionnaires électroniques à des programmes de traitement automatique. Nous trouvons un aperçu des distinctions entre dictionnaires électroniques et dictionnaires informatisés dans l'article de Courtois Blandine et Max Silberztein cité plus haut, mais également dans l'ouvrage du second de ces linguistes, *Dictionnaires électroniques et analyse automatique de textes*. *Le système INTEX*. Paris, Masson, collection Informatique linguistique, paru en 1993 et dans l'article de Maurice Gross, 1989, « La construction de dictionnaires électroniques »[27].

Dans le domaine français de l'élaboration de dictionnaires électroniques, les travaux du Laboratoire d'Automatique Documentaire et Linguistique (LADL) de l'Université de Paris VII, basés sur la description et le recensement du vocabulaire courant, des points de vue syntaxiques, morphologiques et orthographiques, font incontestablement figure de proue et se trouvent illustrés par la mise en place de

[26] In Dictionnaires électroniques du français, *Langue Française, n°* 87, 1990, Paris, Larousse, pp. 11-22.
[27] *Annales des Télécommunications, tome 44, n°1-2, pp. 4-19,* Issy-les-Moulineaux/Lannion : CENT.

plusieurs dictionnaires électroniques, connus sous les noms de DELAS, DELAC, DELAF ou DELAP, sur lesquels nous allons à présent nous attarder quelque peu.

Le Dictionnaire Electronique du LADL pour les mots Simples (DELAS), dont Blandine Courtois sa créatrice fournit une présentation dans diverses communications comme, 1985, « Le système DELAS. Présentation technique»[28], et, 1990, « Un système de dictionnaires électroniques pour les mots simples du français »[29], est une «base de données orthographique et morphologique créée sur support magnétique », pour reprendre les termes de cette dernière, qui comporte environ 80000 entrées regroupant les mots simples du français, c'est-à-dire « des unités de texte définies sur l'alphabet des codes ASCII ou EBCDIC à 256 caractères, et ne comportant aucun séparateur, en particulier pas de trait d'union, ni apostrophe, ni espace blanc » et généralement étant les formes canoniques suivantes : les verbes à l'infinitif, les noms au masculin singulier et les adjectifs ou participes, au masculin singulier également.

La mise en place de cet immense corpus repose sur l'utilisation de divers dictionnaires, essentiellement issus des maisons d'édition Robert et Larousse, fournissant chacun un certain nombre d'acceptions, et qui sont le *Petit Robert* (1982) et le *Grand Robert* (1986), le *Petit Larousse Illustré* (1986), le *Petit Larousse Illustré* (1989), le *Larousse LEXIS* (1979) et le *Grand Dictionnaire Encyclopédique Larousse* (1982).

Le Dictionnaire Electronique du LADL des mots Fléchis (DELAF) est un dictionnaire étroitement lié au DELAS, dont nous trouvons une présentation dans l'article de Blandine Courtois, 1990, « Un système de dictionnaires électroniques pour les mots simples du français »[30], est un dictionnaire directement dérivé du DELAS. Il contient en effet l'ensemble des formes fléchies et conjuguées des mots présents dans celui-ci et ce par le biais d'un programme qui, ainsi que l'illustre Blandine Courtois dans l'article précédemment cité, « effectue la flexion des noms et adjectifs, et la conjugaison des verbes ».

[28] In *Rapport Technique du LADL*, 1985a.
[29] *Langue Française*, n° 87, 1990, Paris : Larousse, pp.11-21.
[30] *Langue Française*, n° 87, 1990, Paris : Larousse, pp.11-21.

A l'image du DELAS qui décrit la morphologie et la flexion des mots simples, le Dictionnaire Electronique du LADL des mots composés (DELAC), décrit celles des mots composés c'est-à-dire des adverbes, des conjonctions de coordination et des noms qui sont orthographiquement composés de plusieurs constituants. Destiné à la reconnaissance automatique des divers mots composés d'un texte, ce dictionnaire repose sur un classement de la structure morpho-syntaxique de son répertoire. Max Silberztein, dans l'article, 1990, « Le dictionnaire électronique des mots composés »[31], fournit une description de ce dictionnaire et de son fonctionnement, et évoque notamment le Dictionnaire électronique du LADL pour les mots Composés Fléchis (DELACF), ouvrage analogue au DELAF, composé d'environ 140000 entrées.

Pour une autre présentation du DELAC, notons également l'article de Blandine Courtois et Max Silberztein, 1989, « Les dictionnaires électroniques DELAS et DELAC »[32].

Le Dictionnaire Electronique du LADL pour les représentations phonémiques (DELAP), est un dictionnaire assez proche du DELAS puisqu'il décrit également, ainsi que l'affirme Eric Laporte dans son article, 1990, « Le dictionnaire phonémique DELAP »[33], « les mots simples ne comportant aucun séparateur, écrits en minuscules sous leur forme canonique ». Il comporte donc lui aussi environ 80 000 entrées, mais associe à celles-ci des informations concernant leur prononciation et leurs variations phonétiques.

Nous trouvons une autre présentation du DELAP, ainsi que de son dérivé le Dictionnaire Electronique du LADL pour les représentations phonémiques fléchies (DELAPF), mais aussi des autres dictionnaires du LADL, dans la thèse de Doctorat de Dominique Revuz, *Dictionnaires et lexiques. Methodes et algorithmes*.

---

[31] *Langue Française* n° 87, 1990, Paris : Larousse, pp. 71-83.
[32] *RELAI : Recherches en Linguistique Appliquée à l'Informatique (Actes du colloque «La description des langues naturelles en vue d'applications informatiques »* (Québec 1988)), Québec : Université Laval.
[33] *Langue Française* n° 87, 1990, Paris : Larousse, pp. 59-70.

Dans ce même ouvrage, se trouve sommairement évoquée la notion de « lexique-grammaire », étroitement liée à celle de « dictionnaire-machine », et qui renvoie à une liste de phrases fournissant la description syntaxique élémentaire des diverses lexies répertoriées dans le DELAS.

Par le biais de ces listes, organisées sous forme de tables, l'utilisateur a ainsi accès aux propriétés distributionnelles et transformationnelles des lexies. A travers l'article de Christian Leclère, 1990, « Organisation du lexique-grammaire des verbes français »[34], nous disposons d'une illustration du fonctionnement d'un lexique-grammaire du LADL. Evoquons également la présentation de la notion de lexique-grammaire dans l'article de Maurice Gross, 1989, « La construction de dictionnaires électroniques »[35].

Face aux dictionnaires mis en place et développés au Laboratoire d'Automatique Documentaire et Linguistique, divers répertoires électroniques ont également été créés en France. Bien loin de prétendre dresser une liste exhaustive de ces derniers, nous pouvons néanmoins signaler les travaux du groupe de recherche PRC (Communication Homme-Machine) au laboratoire IRIT de l'Université Paul Sabatier de Toulouse, qui ont permis la mise en place de deux bases de données connues sous le nom de BDLEX-23000, une base de données contenant environ 23000 entrées sous formes canoniques, et fournissant sur celles-ci, des informations comme la graphie de leur forme accentuée, leur transcription phonologique (indication des frontières de syllabes), les phonogrammes (association lettres/sons), leurs flexions, leur dérivation, leur morphosyntaxe, ou des indices sur leurs fréquences d'apparition dans les textes (le lexique contient environ 270 000 formes fléchies), et BDLEX-50000, une base de données contenant environ 50000 entrées canoniques et fournissant des informations concernant la phonologie, la graphie et la morphosyntaxe (environ 450 000 formes fléchies) de celles-ci.

Il existe aujourd'hui une grande quantité de ce type de dictionnaires électroniques, et pas seulement en France. Nombre d'entre eux sont répertoriés

[34] *Langue Française* n° 87, 1990, Paris : Larousse, pp.112-122.
[35] *Annales des Télécommunications, tome 44, n°1-2, 1989, pp. 4-19*, Issy-les-Moulineaux/Lannion : CENT.

sur Internet, mais à l'image du site suivant, http://www.riofil.org/Publications/Materiaux/Mat_Dict_Elect.html, qui propose un échantillon infime de ces répertoires, la notion de « dictionnaires électroniques » réunie sur ces sites aussi bien des dictionnaires informatisés que des dictionnaires électroniques conçus pour être exploités par des machines.

Qu'il s'agisse des dictionnaires informatisés, c'est-à-dire des répertoires ayant, à l'image du *Trésor de la Langue Française*, subi une rétroconversion, des dictionnaires informatisés élaborés à partir de grands corpus, ainsi que le *Collins COBUILD English Dictionary*, ou des dictionnaires-machines, comme les ouvrages du Laboratoire d'Automatique Documentaire et Linguistique, créés pour être utilisés par des ordinateurs, ces trois types de dictionnaires sont indifféremment regroupés sous l'appellation « Dictionnaires Electroniques ». Or nous venons de voir, en fournissant un aperçu de ces divers produits, qu'il s'agissait en fait d'ouvrages réellement différents les uns des autres.

Malgré leurs disparités respectives, ces derniers possèdent néanmoins de nombreux points communs, dont le plus important réside certainement dans leur nature profonde, celle d'outils apportant un renouveau incontestable à la lexicographie française. Les répertoires électroniques et informatisés ont effectivement contribué depuis leur émergence à modeler un nouveau visage à la lexicographie, en offrant notamment des moyens d'exploitation des données plus performants et relativement simples d'utilisation.

Leur apport, bien qu'appelé à être de plus en plus important, ne semble cependant pas représenter, tout au moins dans un proche avenir, une menace pernicieuse pour l'existence des dictionnaires papiers, mais bien au contraire un véritable enrichissement lexicographique.