HAL
archives-ouvertes.fr

# Le Petit Larousse Illustré de 1905 pris dans la Toile

Helene Manuelian

▶ **To cite this version:**

Helene Manuelian. Le Petit Larousse Illustré de 1905 pris dans la Toile. Cahiers de Lexicologie, Centre National de la Recherche Scientifique, 2006, 1 (88), pp.183-200. <hal-00526599>

**HAL Id: hal-00526599**

**https://hal.archives-ouvertes.fr/hal-00526599**

Submitted on 15 Oct 2010

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Le *Petit Larousse Illustré* de 1905 pris dans la Toile

Hélène Manuélian,
Métadif & Université de Cergy Pontoise.

*Car c'est bien la plus grande vérité de notre ère :l'information n'est pas la connaissance.*

*Caleb Carr (Le tueur de temps, Seuil, 2000).*

Internet est aujourd'hui une source d'information à peine imaginable il y a quelques décennies. Cette technologie du *tout information*, qui consiste à permettre aux Internautes d'accéder à tous les types de données (politiques, artistiques, scientifiques), ne permet pourtant pas toujours de fournir des éléments d'analyse aux lecteurs, à cause du souci d'immédiateté qui parfois fait oublier que la connaissance n'est pas l'information et que le recul est parfois nécessaire à l'analyse et à la compréhension du monde. Ainsi, des technologies nouvelles comme le Wiki dont nous reparlerons plus loin permettent une diffusion massive et immédiate de l'information. On peut donc reprocher à Internet de diffuser de l'information mais de ne plus laisser la place à la connaissance. Aujourd'hui, de nombreuses initiatives se proposent de remédier en partie à ce problème, en proposant des bibliothèques en ligne : on peut citer la bibliothèque Gallica pour la France, et le projet de Google consistant à numériser et à diffuser gratuitement les ouvrages littéraires et scientifiques libres de droit. Cette démarche permet ainsi, tout en diffusant l'information nouvelle, de faire de l'Internet le dépositaire de la culture littéraire et scientifique mondiale, et de donner un accès libre à la connaissance au plus grand nombre.

A l'heure ou les technologies du Wiki se développent sur Internet, l'UMR CNRS / Université de Cergy Pontoise décide d'informatiser le *Petit Larousse Illustré* de 1905. Si cette idée peut paraître étrange et inutile au premier regard, elle se justifie pleinement. Le but de cet article est de faire le point sur les ressources et les connaissances lexicales informatisées à ce jour, et en revenant sur l'histoire du *Petit Larousse Illustré*, de montrer que cette informatisation a un sens, et correspond à un besoin. Nous ferons pour terminer un point sur l'état d'avancement du travail et sur les projets que nous avons dans ce cadre.

## 1. Dictionnaires, encyclopédies et ressources lexicales en ligne

### *1.1. Ressources encyclopédiques et dictionnaires en ligne*

Internet est souvent considéré – à tort – comme une encyclopédie géante. Pour nous, ce n'est absolument pas le cas, ne serait - ce qu'à cause de la méthode de recherche des informations, qui est thématique et ne garantit pas de résultat immédiat. Par ailleurs, le joyeux désordre qui règne sur Internet ne

ressemble pas au contenu d'une encyclopédie. Des sites au contenu peu rigoureux aux sites idéologiquement marqués, des blogs aux chats, tous les types de contenus sont présents sur Internet. Et au milieu de tout cela, les encyclopédies, traditionnelles ou utilisant les technologies du Wiki, et les dictionnaires, résultat du travail de professionnels de la lexicographie, ou d'amateurs. Notre but ici n'est pas de faire un inventaire des ressources lexicographiques existantes sur le Web, mais de faire le tour des types de données que l'on peut trouver.

### *1.1.1. Encyclopédies traditionnelles*

Dans un premier temps, les ressources dictionnairiques les plus connues par le grand public sont les grandes encyclopédies multimédia. Parmi elles, on peut citer l'encyclopédie Encarta, et Webencyclo[1]. Ces encyclopédies présentent les caractéristiques des Encyclopédies papier du point de vue du contenu, avec bien entendu des modalités de consultation liées à l'informatique (possibilités de requêtes, recherches thématiques). Généralement, ce type d'encyclopédie est gratuit, mais leur consultation nécessite une inscription préalable.

### *1.1.2. Encyclopédies en « Wiki »*

Wikipédia[2] est une encyclopédie particulière dans le sens où elle utilise une technologie très moderne, autant du point de vue technique que philosophique : le Wiki. La technologie Wiki permet à n'importe quel internaute de modifier le site qu'il consulte, sans pour autant être un expert en développement de site web. Par le biais de quelques clics de souris, il peut accéder au texte initial du site web et le modifier. Tout un arsenal informatique permet la consultation des versions antérieures du site, celle des modifications ainsi que la signature de l'internaute qui prend l'initiative de le modifier. Ceci dit, la signature de la modification n'est pas obligatoire, et surtout, n'est pas informative la plupart du temps, les utilisateurs ayant tendance à utiliser des pseudonymes.

Pour les lexicographes et encyclopédistes « traditionnels », cette technologie peut paraître excessivement dangereuse. Comment accepter que n'importe qui puisse ajouter, modifier ou supprimer des informations dans un ouvrage considéré comme un ouvrage de référence ? Pourtant, les tenants du Wiki (qui sont souvent aussi des militants du logiciel libre) justifient leur attitude avec des arguments qui méritent d'être écoutés. Ils partent du principe que les personnes qui s'impliquent dans la diffusion de l'information sont de bonne foi, et souhaitent apporter leur savoir dans des domaines sur lesquels ils sont experts. Par ailleurs, ils estiment que la pression du groupe, la vérification des informations par les autres internautes permet la correction quasiment immédiate en cas d'erreur. Ceci qui est encore renforcé par la visibilité

[1] Respectivement localisées aux adresses suivantes : http://fr.**encarta**.msn.com et http://www.**webencyclo**.com/home.asp

[2] Localisée à l'adresse http://fr.wikipedia.org/wiki/France

permanente de l'historique des changements opérés. On a pu alors constater au soir même de la nomination de D. de Villepin comme Premier ministre la modification de l'article le concernant.

Si nous nous attardons sur cette technologie, c'est parce que nous pensons qu'il s'agit d'une technologie et d'une philosophie en pleine expansion, qu'elle a un avenir réel dans le monde de la lexicographie multimédia, et qu'elle va probablement changer la vision de l'encyclopédie de façon générale. Il nous semble que la lexicographie moderne devra en tenir compte, et l'intégrer dans ses réflexions d'une manière ou d'une autre, tout en gardant à l'esprit qu'il est probablement dangereux de laisser l'actualité et la simple information prendre le dessus sur la connaissance et l'analyse.

### *1.1.3. Dictionnaires professionnels*

On trouve aussi sur Internet de nombreux dictionnaires réalisés par des professionnels. Ainsi, on peut citer le monument de la lexicographie informatisée en France, le *Trésor de la Langue Française Informatisé*[3], réalisé par le CNRS (laboratoire ATILF). Ce dictionnaire, écrit par des lexicographes professionnels, propose aujourd'hui des modalités d'interrogation inégalées à présent, avec des possibilités de requêtes extrêmement complexes, ainsi que la possibilité d'hypernavigation avec la base Frantext et d'autres dictionnaires ou bases de données lexicales. On peut aussi consulter la huitième et la neuvième édition du dictionnaire de l'Académie.

On peut citer ici aussi le projet Papillon (Mangeot-Lerebours et al., 2003) qui utilise une technologie proche du Wiki pour créer un dictionnaire français/japonais accessible aux locuteurs de deux langues (qui en général ont un problème lié à la différence de système d'écriture des deux langues).

Ces dictionnaires réalisés par des professionnels (de la linguistique en général, de la lexicographie et de l'informatique) ont l'avantage de constituer des ressources fiables, dont la maintenance est assurée (surtout quand ils sont réalisés par des institutions publiques).

### *1.1.4. Dictionnaires amateurs*

Nous venons de parler des avantages des ressources linguistiques créés par des institutions, mais nous ne pouvons pas – étant donné la nature de l'Internet – ne pas avoir un regard sur les sites de dictionnaires réalisés par des amateurs. Certes, il ne présentent pas le degré de fiabilité qu'on attend des dictionnaires institutionnels. Pourtant, leurs rédacteurs sont souvent des passionnés, experts de leurs domaines, et leur existence permet aussi de donner une vitalité inégalable à ce type de ressources. Ainsi, on trouve à l'heure actuelle un dictionnaire de néologismes, récoltés par son auteur dans le Monde et le Soir

[3] http://www.atilf.fr

de Bruxelles, un dictionnaire de la Zone[4], qui décrit le langage des banlieues, et bien d'autres encore. Il nous semble évident que ces dictionnaires n'étant réalisés en général que par une seule personne, n'ayant pas forcément de connaissances très approfondies sur la langue, ils sont à manipuler avec distance et attention. Cependant, ils constituent des ressources intéressantes plus en terme d'information qu'en termes de connaissance, contrairement aux dictionnaires institutionnels.

### *1.2. Dictionnaires anciens en ligne*

La démocratisation de l'Internet a aussi permis a des chercheurs de rééditer des dictionnaires du passé, dont les exemplaires devenaient rares et chers. Ainsi, le projet ARTFL de Chicago[5] et l'ATILF proposent de consulter des dictionnaires anciens : le dictionnaire critique de Féraud, les anciennes éditions du dictionnaire de l'Académie, le Thrésor de Nicot. Ces sites Internet, tous en accès gratuit permettent aujourd'hui aux chercheurs comme aux amateurs de consulter à loisir ces monuments du patrimoine lexicographique francophone.

### *1.3. Dictionnaires électroniques*

La dernière catégorie de dictionnaires qu'on retrouve en ligne sont les dictionnaires électroniques. Ces dictionnaires n'ont pas du tout les mêmes fins que les autres dictionnaires informatisés. Le terme dictionnaire électronique a été employé en France pour la première fois par Maurice Gross (1975). Il s'agit de dictionnaires destinés aux machines, pour leur faire faire du traitement automatique des langues. Ces dictionnaires servent de base de données pour l'analyse morphologique, syntaxique, et sémantique des textes. On pourra citer des ressources comme FLEMM ou les données issues du projet MorTAL (Dal et al. 2004) pour la morphologie, les dictionnaires destinés à la syntaxe du LADL (Gross, 1991) et du LLI[6], ainsi que les bases de données sémantiques (n'existant que dans une très faible mesure pour le français) de type Wordnet (exprimant des relations lexicales, (Fellbaum, 1998)) ou Framenet (exprimant des restrictions de sélection pour les verbes (Baker et al. 1998)).

Toutes les données que nous venons de décrire forment un ensemble cohérent, riche et en accès libre qu'il est donc important de connaître et d'utiliser. En effet, nous disposons aujourd'hui d'une véritable bibliothèque en ligne, offrant des modalités de consultation intéressante et une puissance de recherche bien supérieure à celle qu'offrait une bibliothèque papier, même si elle ne nous dispense pas du travail d'analyse réalisé par un humain. A ces outils d'information, doivent en effet s'ajouter des analyses, des travaux, et une

---

[4] Consultables aux adresses suivantes : http://membres.lycos.fr/antidico/ et http://cobra.le.cynique.free.fr/dictionnaire/

[5] http://humanities.uchicago.edu/orgs/ARTFL/

[6] http://www-lli.univ-paris13.fr/

connaissance de la langue. Aujourd'hui, le laboratoire Métadif propose d'informatiser la première édition du *Petit Larousse Illustré*, celle de 1905. Cet outil informatisé viendra s'ajouter aux ressources existantes, et il maintenant nécessaire de motiver ce projet, et de montrer en quoi il ne constituera pas seulement de l'information, mais un outil supplémentaire d'accès à la connaissance. Pour motiver notre travail, nous allons dans un premier temps revenir sur son histoire, et ses caractéristiques.

## 2. Histoire du *Petit Larousse Illustré* et de cent ans de succès

### *2.1. Naissance en 1905*

Le *Petit Larousse Illustré* naît en 1905, trente ans après la disparition de Pierre Larousse, co-fondateur de la maison du même nom. Sa première édition, dirigée par Claude Augé, est mue par les mêmes idéaux que ceux qui ont permis de créer les autres dictionnaires Larousse. Créée en 1856, la librairie Larousse est le fruit de la réflexion de deux instituteurs, Pierre Larousse et Augustin Boyer, amoureux de la langue française et défenseurs de la République. De façon générale, les éditions Larousse affichent leur volonté de diffusion du savoir en permettant l'accès à la langue et à l'orthographe françaises. Le *Petit Larousse Illustré* s'inscrit directement dans ce mode de pensée. En créant ce « petit » dictionnaire, Claude Augé offre au plus grand nombre un dictionnaire compact, attrayant grâce à ses illustrations et offrant la possibilité de connaître l'orthographe, la prononciation et le sens des mots du français. Par ailleurs dictionnaire encyclopédique, le *Petit Larousse Illustré* de 1905 présente aussi une partie réservée aux noms propres, et les désormais célèbres pages roses de locutions latines. Claude Augé tient au caractère encyclopédique de son dictionnaire, et ainsi Jean Pruvost (2004) analyse la première page du dictionnaire de la façon suivante : « La langue demeure prioritaire, même s'il faut, pour la décrire dans toute sa dimension, ne surtout pas la priver des référents du monde et donner aux mots, à travers les exemples, toute leur force encyclopédique »[7]. Cette caractéristique nécessitera de nombreuses refontes au cours du siècle, et nous allons maintenant les décrire.

### *2.2. Des millésimes aux refontes*

L'une des caractéristiques du *Petit Larousse Illustré* est donc d'être ré-édité tous les ans, et de subir des refontes régulièrement au cours de son siècle d'existence. Ce dictionnaire devient donc un dictionnaire totalement inscrit dans son époque, et reflétant au travers de son histoire, les évolutions sociales importantes du XX ème siècle.

La réédition annuelle et les multiples refontes ont permis de faire de l'ensemble des *Petit Larousse Illustré* un magnifique corpus du français du XXème siècle. Les mots apparaissent, disparaissent, les définitions évoluent en

[7] In *La dent de lion, la semeuse et* le Petit Larousse, p.57

fonction des tabous (on notera un travail de maîtrise réalisé récemment par J. Masmejean (2005) étudiant les définitions des termes liés à l'érotisme et la sexualité et démontrant que le dictionnaire est un miroir de la société qui le produit). Par ailleurs, dans la partie des noms propres, on voit apparaître des personnages régulièrement, dans le domaine artistique ou politique.

### *2.3. Des illustrations alliées à la simplicité*

#### *2.3.1. Un dictionnaire pour tous*

Deux éléments font partie des raisons du succès du Petit Larousse. Son format en fait un dictionnaire pour tous. Bien qu'encyclopédique, ce dictionnaire s'impose des définitions simples et courtes, ce qui le rend très grand public. Pour appuyer les définitions, les illustrations sont présentes à toutes les pages (5800 gravures, 130 tableaux et 120 cartes sont annoncés pour la première édition, 5000 illustrations et 321 cartes pour l'édition du centenaire). Ces deux éléments combinés en font un réel outil de transmission du savoir, comme en témoignent les écrits d'Azouz Begag et de Bernard Pivot, (cités par J. Pruvost, 2004) tout deux nostalgiques à l'évocation de ce dictionnaire, et pourtant de générations et d'origines sociales différentes.

#### *2.3.2. Variété des illustrations*

Les illustrations du Petit Larousse sont extrêmement riches, et on en trouve différents types.

*Vignettes, Lettrines et Culs de Lampes :* Ces trois éléments apparaissent dès la première édition du *Petit Larousse Illustré*. Ils seront parfois supprimés au cours des différentes refontes, mais font partie de son identité. Les vignettes capitulaires sont particulièrement célèbres. Apparaissant à chaque chapitre, elles correspondent à une lettre (qui apparaît au milieu), et illustrent la lettre en question en représentant des animaux, des objets, des personnages et des éléments naturels dont le nom commence par la lettre correspondant au chapitre.

*Les planches* : Comme les vignettes capitulaires, les planches sont des illustrations dont on ne se lasse pas d'étudier les détails. Elles présentent les diverses espèces d'une race animale, d'un végétal, et on se régalera de constater leur évolution au cours du temps. Gravures, photos, dessins, en couleur ou en noir et blanc, toutes les méthodes d'illustration ont été utilisées, ont fait du Petit Larousse ce dictionnaire qu'on peut à tout âge, consulter en laissant son imagination vagabonder au fil des pages.

*Les exemples* : Bien entendu, l'illustration c'est aussi l'illustration linguistique. Le *Petit Larousse Illustré* regorge d'exemples, puisque selon Pierre Larousse lui même, « un dictionnaire sans exemples est un squelette ». La plupart du temps, il s'agit d'exemples forgés, qui permettent au lecteur d'être éclairé à la fois sur le sens et l'emploi du mot, mais parfois aussi sur les aspects

encyclopédiques de l'objet auquel il réfère. D'après J. Pruvost, on trouve quelques citations, essentiellement dans la partie noms propres du dictionnaire.

Le Petit Larousse est toujours, cent ans plus tard, un immense succès éditorial. Sa vision démocratique de la transmission du savoir, sa mise à jour annuelle et ses illustrations amusantes en font le plus célèbre des petits dictionnaires. Aujourd'hui confronté à de nombreux concurrents de qualité, il reste le préféré, grâce à une ligne éditoriale constante, en adéquation avec les évolutions sociales, mais aussi parce qu'il est le plus ancien, celui que les lecteurs connaissent depuis toujours.

### *2.4. A l'heure des technologies du Wiki, informatiser le Petit Larousse de 1905… Quelle drôle d'idée !*

Nous venons de le voir, la dimension affective du Petit Larousse comme le dictionnaire de l'enfance est immense. Loin de nous l'idée de pouvoir transposer le plaisir de tourner ses pages jaunies à l'écran, nous pensons cependant qu'il est important d'informatiser et de diffuser ses éditions anciennes.

#### *2.4.1. Souci démocratique de Pierre Larousse*

Tout d'abord, nous souhaitons inscrire notre démarche dans la philosophie Laroussienne. En effet, nous savons que Larousse était un instituteur, républicain, et que l'un de ses soucis était de faciliter l'accès à la connaissance par l'intermédiaire de l'enseignement du Français. Aujourd'hui, nous pensons que l'accès à la connaissance de la langue française peut passer par une version informatisée des dictionnaires grands public. En effet, s'il existe déjà de nombreux dictionnaires informatisés, ce ne sont pas forcément ceux que connaît le grand public. En dehors des dictionnaires anciens ou des dictionnaires encyclopédique, à notre connaissance, seul le TLFI peut être cité comme dictionnaire monolingue de langue en ligne et gratuit. Si ce dictionnaire est – nous ne le répéterons jamais assez – un outil inégalé et un monument lexicographique, il est relativement peu connu du grand public. Certes le *Petit Larousse Illustré* est un dictionnaire qui peut paraître désuet, mais il peut aussi constituer un premier pas dans la direction de l'informatisation de dictionnaires destinés au grand public, monolingues, de langue.

#### *2.4.2. Le Petit Larousse est un patrimoine linguistique, et par là, un reflet de la société française*

Nous pensons aussi qu'informatiser le Petit Larousse de 1905 et les éditions suivantes si cela est possible, offrira aux chercheurs un outil extrêmement utile, aussi bien pour des linguistes, historiens, ou sociologues. En effet, pour les linguistes, il constitue un très riche corpus de par ses exemples (la plupart du temps forgés) qui viendra compléter les corpus littéraires, en donnant une illustration de l'utilisation standard du lexique de l'époque. Ainsi,

l'ensemble des *Petit Larousse Illustré* peut être considéré comme une série d'instantanés du français du XX ème siècle.

Pour les chercheurs en sciences humaines de façon générale, il présente un intérêt historique indéniable, puisque comme tous les dictionnaires, il est le reflet de l'idéologie, des connaissances et plus généralement de la société qui lui est contemporaine.

#### *2.4.3. Intérêt métalexicographique*

Plus précisément, l'informatisation du Petit Larousse de 1905, et encore plus si elle est suivie de l'informatisation des éditions lui succédant, présentera un intérêt en métalexicographie. En effet, on pourra accéder à une présentation synoptique des définitions et ainsi, plus facilement observer les évolutions dans la rédaction et le contenu des définitions, dans le choix des exemples les accompagnant, et dans tous les autres éléments qu'un métalexicographe peut vouloir observer.

#### *2.4.4. Intérêt en lexicologie informatique*

Le dernier intérêt que nous voyons à l'informatisation du *Petit Larousse Illustré* est un intérêt à plus long terme. Comment ne pas envisager, quand on appartient au domaine du traitement automatique de la langue, d'en faire une base de données sémantique ? L'idée n'est pas neuve, mais a finalement été assez peu réalisée pour le français. Ainsi, une fois porté sur support numérique, on peut envisager de pousser l'informatisation jusqu'à en faire une base de donnée formalisée, représentant les données sémantiques aussi bien en termes de relations lexicales, actancielles ou même des relations plus lâches. Ainsi, ce souhait d'utiliser les définitions de dictionnaires et non les descriptions formelles du sens pour résoudre des problèmes de traitement automatique des langues comme la résolution ou la génération de reprises coréférentielles et d'anaphores associatives a été exprimé par Gardent, Manuélian et Kow (2003), et Manuélian (2003).

On nous objectera que la base de données que constitue le Petit Larousse de 1905 est désuète, mais nous pensons qu'à défaut de constituer une ressources définitive, elle présente deux intérêts : celui d'être assez petite pour servir d'expérimentation sans le déploiement d'énergie et de moyens considérables nécessité par un dictionnaire plus important en taille ; l'autre de pouvoir être utilisable pour tester la résolution automatique des reprises et anaphores dans des textes déjà anciens, ce qui pourrait présenter un intérêt dans le cas d'études diachroniques.

## 3. L'informatisation du millésime 1905 : Etat des lieux et projets

Après avoir présenté l'intérêt de l'informatisation d'un dictionnaire comme le *Petit Larousse Illustré*, nous nous devons maintenant de faire le point sur le projet, tel qu'il se présente à l'heure où nous écrivons. Le travail

d'informatisation d'un dictionnaire se présente toujours de la façon suivante. Tout d'abord, il faut récupérer le texte sur un support numérique ; ensuite, le texte doit être balisé de façon à ce que les informaticiens puissent donner la possibilité aux utilisateurs de faire des requêtes sur le texte. Nous présenterons dans un premier temps le travail réalisé, puis nous présenterons le travail en cours et à venir.

## *3.1. Le travail réalisé*[8]

### *3.1.1. Numérisation*

La transposition du support papier au support numérique n'est pas toujours une manipulation facile, surtout lorsque les ouvrages sont anciens. Lorsqu'ils sont trop fragiles, une saisie manuelle du texte est obligatoire, et demande un investissement en temps et en personnel très important. Pour le *Petit Larousse Illustré* de 1905, il nous a été possible de scanner le texte. En effet, il existe encore de nombreux ouvrages, en relativement bon état.[9] L'équipe de numérisation a donc fait appel à une société privée spécialisée dans la numérisation, et qui a réalisé le travail en trois jours, fournissant ainsi à notre laboratoire, deux types de fichiers : un fichier image (résultat du scanner), et un fichier texte, résultat du processus de reconnaissance de caractères. En effet, pour pouvoir informatiser le dictionnaire (nous y revenons au paragraphe suivant), il nous faut absolument récupérer le texte, pouvoir travailler le texte lui-même, et non pas simplement une photographie du texte (ce qu'est le fichier image).

Une phase de relecture et de nettoyage du texte sera alors nécessaire car le résultat de la reconnaissance de caractère n'est jamais parfait.

### *3.1.2. Nettoyage du texte numérisé*

La relecture permet de constater que la reconnaissance de caractère n'est juste qu'à 50%. Après un an de travail (deux personnes, à mi-temps), environ la moitié du texte a été relue (nous ne traitons pas la partie noms propres pour le moment). Il est prévu de faire trois relectures de façon à laisser le moins possible d'erreurs dans le texte. Nous pensons pouvoir terminer le travail pendant le premier trimestre 2006.

### *3.1.3. Analyse lexicographique du texte*

Jean Pruvost fournit à l'équipe une analyse lexicographique minutieuse du *Petit Larousse Illustré* de 1905, de façon à permettre le balisage du texte qui

---

[8] Nous résumons ici le travail. Pour un exposé plus détaillé des phases de numérisation et de relecture, nous renvoyons à Manuélian et Timmermann (à paraître)

[9] Les éditions Larousse ont réédité à l'occasion du centenaire du *Petit Larousse illustré* le fac-similé de l'édition de 1905, malheureusement la numérisation s'est faite un an auparavant sur un ouvrage de l'époque.

en permettra la consultation électronique. De l'analyse de Jean Pruvost, se dégage tout un tableau de marqueurs lexicographiques qui concernent tant la macro- que la microstructure. Cette analyse permettra un balisage très précis qui devrait faciliter les requêtes des lexicographes.

#### *3.1.4. Prébalisage*

Cinquante pages du *Petit Larousse Illustré* de 1905 sont alors prébalisées, de façon à tester la possibilité de baliser un niveau très fin d'analyse lexicographique. La série de prébalises créée pour l'occasion était totalement ad hoc, et bien entendu, nous envisageons d'utiliser maintenant les balises correspondantes utilisé dans les grands projets de normalisation des ressources textuelles (TEI et comités de l'ISO).

### *3.2. Résultats souhaités et outils utilisés*

Afin de motiver les travaux que nous menons actuellement sur le balisage, nous souhaitons maintenant présenter précisément les buts que nous nous fixons.

#### *3.2.1. Une ressource en accès libre*

Bien qu'aujourd'hui le *fac simile* de l'édition 1905 soit à nouveau en vente, nous souhaitons en faire une ressource libre d'accès sur Internet En effet, nous pensons qu'il s'agit d'un outil complémentaire des autres dictionnaires déjà en ligne gratuitement, et qu'il serait dommage de ne pas le diffuser largement.[10]

#### *3.2.2. Requêtes sur des types d'objets*

Pour nous, l'informatisation d'un dictionnaire ancien ne présente pas d'intérêt si elle n'apporte pas plus de possibilités que la consultation du dictionnaire papier. Aussi pour nous, l'informatisation ne se limite pas à la conversion des fichiers issus de la reconnaissance de caractères en fichiers HTML par exemple (fichiers mis en forme de façon à être lisibles sur un navigateur web, comme simulé sur la figure 1)

[10] On peut aujourd'hui dire que la mise en ligne ne va pas à l'encontre des intérêts commerciaux : l'exemple du TLFI, en accès libre sur le web, n'a pas empêché la vente d'un CD-Rom.

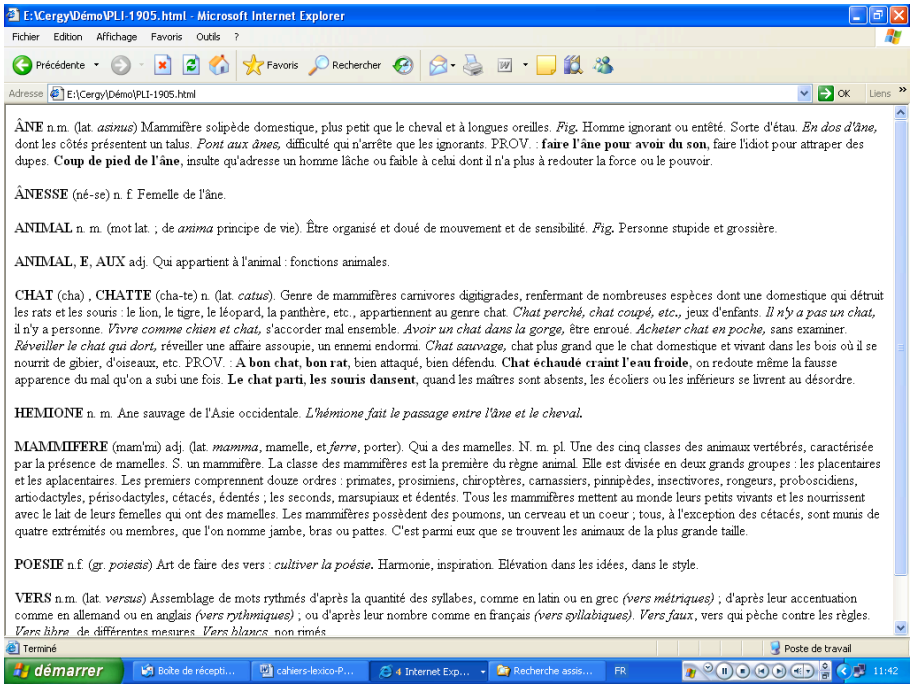


*Figure 1 : Une simple reproduction du texte au format HTML*

Sans avoir l'ambition de créer un outil aussi puissant et aussi performant que le Trésor de la Langue Française Informatisé – ce qui de toutes façons ne serait pas possible, puisque nous ne disposons pas de toutes les ressources textuelles reliées à ce dictionnaire – nous souhaitons rendre possible la formulation de requêtes d'une complexité variable :

Nous souhaitons permettre la simple recherche de définition, où la question sera simplement « Je cherche la définition associée à la vedette « hémione » » mais aussi des recherches plus complexe du type « je cherche tous les mots contenant le nom « âne » dans leur définition (et donc pas dans les exemples) ». C'est à cela que servira le balisage issu de l'analyse lexicographique fournie par Jean Pruvost. En effet, pour permettre une requête, il est nécessaire de baliser le texte pour indiquer à la machine le type d'objet dans lequel elle doit rechercher l'information (définition, exemple, étymologie, etc.).

### *3.2.3. Hypernavigation*

Nous souhaitons par ailleurs permettre à l'utilisateur d'accéder par un simple clic à la définition des termes formant le texte d'autres définitions. Ainsi, comme le montre la la figure 2, la consultation de la définition de « ânesse » permet d'accéder à la définition de « âne », grâce au lien hypertexte inséré sur le mot « âne » contenu dans la définition du nom «ânesse », et au mot « mammifère » grâce au lien inséré dans la définition de « âne »..

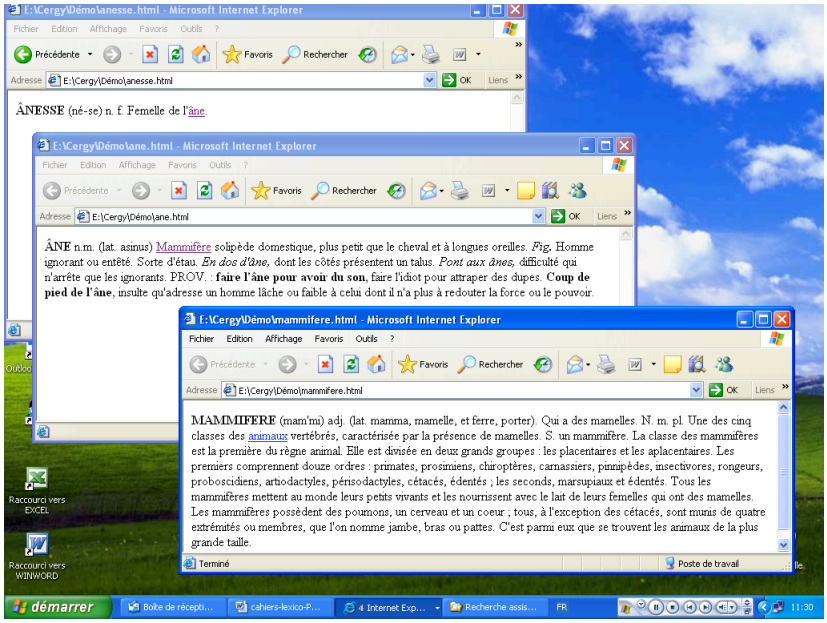


*Figure 2 : Résultat de l'insertion de liens hypertextes*

## *3.3. Outils utilisés*

### *3.3.1. Outils de balisage standard : XML, XSL, HTML*

*Nécessité de tenir compte des normes et standards informatiques* : Notre volonté de diffusion et de partage des ressource a une conséquence directe en informatique : il est nécessaire que les formats que nous utilisons soient des formats standards (par opposition aux formats qu'on appelle propriétaires, qui ne sont utilisables qu'avec un seul système d'exploitation, ou une seule application, et qui nécessitent généralement l'achat de logiciels ou de machines onéreux.). Par ailleurs, l'intérêt d'utiliser des formats standards réside dans la volonté de conservation du patrimoine que nous affichons. En utilisant ces formats, nous avons la garantie qu'au fur et à mesure de l'évolution des technologies, nos fichiers seront lisibles par les nouveaux logiciels, ou en tous cas, possible à convertir pour être lus par les futurs logiciels. Ceci n'est absolument jamais garanti par les logiciels et les formats commerciaux. Nous allons donc maintenant détailler les normes et les standards que nous allons utiliser.

*Standards de codage des fichiers :* Le premier élément important est le format du fichier. Il doit pouvoir être lu sans difficulté par n'importe quelle machine, et prendre le moins de place possible en volume, d'où nos choix pour XML, HTML et tous les formats recommandés par les comités de normalisation des ressources textuelles.

- **XML** : Le premier élément important pour l'informatisation d'une ressource textuelle est de la convertir dans un format de fichier standard. Pour la rendre lisible et exploitable, il est nécessaire que le texte soit balisé, c'est à dire qu'on y ait inséré des éléments correspondant à des indications sur le contenu du texte, que la machine puisse interpréter. Le format standard pour les balises est le format XML, qui n'est en fait pas un langage de balisage à proprement parler, mais un protocole de stockage et de gestion de l'information.

- **XSL et HTML :** HTML est le format standard d'un navigateur web. Il est donc nécessaire pour une consultation libre de notre dictionnaire en ligne de l'utiliser. XSL est un langage qui permet de transformer un fichier XML en fichier HTML, lisible sur le Web.

- **Modèle de document :** XML fournit ce qu'on appelle un *modèle de document,* qui est un ensemble de règles propres à un type de document (roman, dictionnaire, article de presse, etc.). Ces règles permettent de comparer le document produit à un document du même type et de dire s'il est conforme aux règles. On parlera alors de validation du document par le modèle. La plupart du temps, un modèle de document est ce qu'on appelle une DTD (Document Type Definition), mais on trouve aujourd'hui ce qu'on appelle des schémas XML. La DTD est un ensemble de règles qui indiquent quelles balises le document peut utiliser en fonction de sa nature. Elle fournit une description formelle de l'organisation de l'information au sein du document, la liste des attributs possibles pour une balise et les valeurs possibles de ces attributs. On fait référence à la DTD utilisée au début du document pour que XML puisse valider le document. Les DTD peuvent être normalisées si on se réfère aux recommandations de la TEI.

- **La TEI (Text Encoding Initiative) et les recommandations du comité de l'ISO TC37/SC4 :** La TEI est un projet international mis en place à la fin des années quatre-vingts dans le but de créer un environnement dans lequel les documents pourraient être encodés de façon à ce que leurs propriétés soient transcrites et que leur transcription puisse être échangée et survivre aux évolutions technologiques (Mueller, 2002, ISO). Concernant les dictionnaires, tout un chapitre de recommandation a été rédigé au sein du consortium TEI : il s'agit du chapitre 12 : *Print Dictionaries* (Sperberg-McQueen, Burnard, 2004). Parallèlement à la TEI qui est constituée d'experts du domaine affichant la volonté de normaliser les ressources textuelles, mais qui reste une initiative privée, on trouve un des sous comités de l'ISO (International Standard Organisation), le sous – comité 4 du comité technique 37 (désormais TC37/SC4), dont la fonction est à un niveau plus officiel et tout à fait aussi international, de valider les propositions de normalisation des ressources textuelles, et de publier des recommandations. Bien entendu, très vite, le TC37/SC4 a intégré les recommandations de la TEI dans ses normes.[11]

[11] http://www.tc37sc4.org

*Standards de structuration des données :* Actuellement, au travers du projet LMF (Lexical Markup Framework), les comités de l'ISO ouvrent un projet de spécification de structure de bases de données lexicales et lexicographiques. La norme LMF aura pour but de produire des formats standards pour tous les types de bases lexicales, dont les dictionnaires. Elle s'appuie sur les travaux menés dans la TEI et constituera une base de réflexion sur la façon de structurer les dictionnaires pour la prochaine version de la TEI (P5).

De façon générale, l'intérêt d'utiliser les normes existantes est double : il permet de construire la nouvelle ressource grâce à des formats d'échanges des données simples à utiliser, et il permet au reste de la communauté scientifique d'accéder aux données du Petit Larousse sans problèmes techniques. En effet, le coût énorme de l'informatisation (création et maintenance) de telles données nous poussent à dire qu'elles ne doivent pas être construites dans l'isolement, mais reliées à d'autres initiatives, de façon à bénéficier d'un enrichissement mutuel. Notre but sera donc de toujours viser une compatibilité totale du point de vue informatique, aussi bien en termes de logiciels que de systèmes d'exploitation

#### *3.3.2. Logiciels de traitement automatique des langues*

Afin de réaliser le balisage lexicographique du texte du *Petit Larousse Illustré* (nous le montrerons plus loin), nous allons avoir besoin de logiciels de traitement automatique des langues. En effet, il est difficile de faire autrement si nous souhaitons baliser le texte automatiquement. Nous aurons donc besoin de logiciels d'étiquetage et d'analyse morphosyntaxique. Notre choix n'est actuellement pas arrêté sur un logiciel précis, il devra simplement être compatible avec notre choix de format (c'est à dire accepter de traiter des fichiers au format XML et produire des fichiers XML en sortie).

### *3.4. Travaux en cours*

Actuellement, nous travaillons parallèlement au nettoyage des fichiers et à l'automatisation du balisage du dictionnaire. De nombreux problèmes se posent pour cette automatisation, et nous les présentons ici, ainsi que les solutions envisagées pour les résoudre.

L'automatisation du balisage est nécessaire : manuellement, il est impossible de baliser plus de dix entrées du dictionnaire par jour, et le dictionnaire en comporte environ quarante mille. Par ailleurs, la réalisation manuelle d'un tel travail nous expose à un risque d'erreur important. L'automatisation se heurte à deux types de problèmes que nous développons maintenant : le premier problème est l'automatisation du balisage des éléments constituant les définitions et qui permettra de faire des requêtes dans le dictionnaire, le second est l'automatisation de l'insertion de liens hypertextes entre les définitions.

### *3.4.1. Balisage du contenu des définitions*

Le premier problème auquel nous devons faire face est le problème du balisage des définitions au niveau métalexicographique. Pour baliser le texte automatiquement à ce niveau, il faudrait que la machine puisse reconnaître, sur la base d'indices fiables, les éléments composant la définition Ceci semble impossible, en raison des éléments suivants :

*Non homogénéité de la rédaction* On trouve par exemple les différences suivantes : il existe deux entrées pour le mot *animal*, une pour l'adjectif, une pour le nom, alors qu'il n'y a qu'une seule entrée pour les deux catégories grammaticales du mot *mammifère*. Nous trouvons deux entrées distinctes pour *âne* et *ânesse*, alors qu'il n'y a qu'une seule entrée pour *chat* et *chatte*. Nous ne pouvons alors pas faire en sorte de créer un programme qui considèrerait qu'il n'y a qu'une seule indication de catégorie grammaticale par entrée, ou encore un programme qui n'insérerait qu'une seule balise pour le genre des noms.

*Des marques typographiques identiques pour des informations différentes* : On pourrait alors imaginer de se baser sur la typographie pour repérer certaines informations (ceci nécessiterait alors de conserver la mise en forme après l'OCR, ce qui n'est pas évident, mais possible). Cependant, nous observons que les marques typographiques ne sont jamais univoques. On ne peut donc pas espérer récupérer les informations sur la typographie pour permettre la reconnaissance de certains éléments de contenu.

*L'absence d'indications formelles pour le passage d'une information à une autre* : L'article *mammifère*, par exemple, comporte une définition pour l'adjectif et une définition pour le nom. Les deux définitions sont écrites dans un seul et même paragraphe. Le saut de ligne qui aurait pu aider à délimiter les deux définitions n'était pas présent, on n'a donc encore une fois pas d'indication formelle nous permettant de délimiter des éléments de contenu.

### *3.4.2. Insertion des liens hypertextes*

Il est nécessaire de réaliser le balisage automatiquement, autant pour les requêtes que pour ajouter des liens hypertextes entre les définitions qui permettraient une navigation plus complète.

Par exemple, nous souhaitons qu'en accédant à la définition de *ânesse* (*femelle de l'âne*), l'utilisateur puisse ensuite directement lire la définition de *âne*. Nous pensons pour l'instant à ne proposer des liens que sur les termes représentant les classificateurs permettant la définition, mais cela pose déjà un certain nombre de problèmes. Nous souhaitons pouvoir insérer automatiquement les liens dans le texte, pour des raisons de temps, ce qui nous amènera à insérer des balises que nous appellerons source du lien – à l'intérieur des définitions, et des balises cibles – sur la vedette sur laquelle le lien doit pointer. Les problèmes que nous allons rencontrer seront les suivants :

*La reconnaissance automatique des classificateurs (balises sources)* : Il est impossible pour reconnaître le classificateur, d'utiliser la forme de la définition. Très souvent, il est le premier mot de la définition, mais ce n'est pas toujours le cas (cf. la définition de *ânesse* citée précédemment).

Par ailleurs, nous ne pouvons pas envisager d'utiliser la forme des mots contenus dans la définition. En effet, la forme du classificateur peut varier : il arrive que les formes soient fléchies dans les définitions, ce qui ne sera bien entendu pas le cas dans les vedettes. Ainsi, la définition du mot *mammifère* contient la forme *animaux* et non animal, ce qui rend impossible la réalisation d'un programme informatique basé sur la reconnaissance des formes.

*Le problème de l'identification de la cible (vedette sur laquelle le lien doit pointer)* : Enfin, reconnaître la vedette cible du lien n'est pas directement possible. Ici se pose le problème des homonymes. Pour la définition de *mammifère*, par exemple, nous souhaitons pointer vers le nom *animal*, qui est le classificateur utilisé dans la définition. Etant donné qu'il existe une entrée différente pour le nom et pour l'adjectif *animal*, un programme basé sur la reconnaissance de lemmes ne sera pas suffisant.

Les problèmes que nous allons rencontrer vont nécessiter pour automatiser le balisage, et en particulier l'insertion de liens hypertextes, une analyse morphosyntaxique du texte avec reconnaissance des lemmes, de manière à pouvoir trouver dans le balisage une information sur les catégories grammaticales et sur la forme de l'entrée quand elle n'est pas fléchie. Ces éléments vont nous permettre d'effectuer une autre forme de prébalisage (on peut parler de couche préliminaire de balisage), au niveau morphosyntaxique, de façon à pouvoir appliquer ensuite des programmes qui baliseront dans un deuxième temps le texte au niveau lexicographique.

### *3.4.3. Traitement des illustrations*

Comme nous l'avons souligné au début de notre article, une grande partie du charme du *Petit Larousse Illustré* est lié à ses illustrations. Nous souhaitons bien entendu les faire figurer dans la version électronique. Elles sont à l'heure actuelle numérisées, posent des problème spécifiques (particulièrement le traitement des légendes et des textes qui sont insérés à l'intérieur), mais nous avons choisi de les traiter après le texte. Le minimum à faire sera de fournir des liens sur les illustrations correspondant aux définitions, et l'idéal sera de parvenir à traiter le texte et les légendes qui les accompagnent, de façon à les faire apparaître de façon indépendante des définitions lors des réponses aux requêtes des utilisateurs.

## Conclusion

Nous avons montré dans cet article que le *Petit Larousse Illustré* prend toute sa place dans les ressources disponibles sur Internet aujourd'hui. Son informatisation s'inscrit dans un cadre de diffusion des connaissances massive.

Son histoire, ses caractéristiques en font un outil complémentaire aux ressources existantes, et permettrait sa conservation.

Nous avons ensuite exposé l'état d'avancement de ce projet qui n'en est qu'à ses débuts, et les projets que nous avons. Ces projets vont poser un certain nombre de problèmes techniques et théoriques pour lesquels nous n'avons pas encore de réponse, mais nous savons que certains y ont déjà répondu (en particulier lors de l'informatisation du TLF, (Dendien et Pierrel 2001)). Nous allons bien sûr utiliser ces réponses, et essayer pour notre part d'y apporter si c'est nécessaire, des solutions plus simples ou innovantes, dans la mesure où les outils informatiques se sont développés depuis les premières informatisations de dictionnaires.

## Bibliographie


AUGE Claude (sous la dir. de) (1906), *Petit Larousse Illustré*, Paris : éditions Larousse.

BAKER C.F., FILLMORE C.J., LOWE J.B. (1998), The Berkeley Framenet Project in *Proceedings of the thirty-sixth Annual Meeting of the ACL and Seventeenth International Conference on Computational Linguistics*

BONHOMME Patrice (2000), Codage et normalisation de ressources textuelles, *in Ingénierie des Langues*, sous la direction de J-M Pierrel, Hermès, Paris.

DAL Georgette, HATHOUT Nabil et NAMER Fiammetta (2004), Morphologie constructionnelle et Traitement Automatique des Langues : Le projet MorTAL , Lexique n°16 : La formation des mots : horizons actuels, sous la direction de D. Corbin, P. Corbin et M. Temple, Presses Universitaires du Septentrion

DENDIEN Jacques, PIERREL Jean-Marie (2003) Le Trésor de la Langue Française Informatisé. Un exemple d'informatisation d'un dictionnaire de langue de référence, Traitement Automatique des Langues, Volume 44, n°2/2003, Paris, Hermès.

FELLBAUM C.,(1998) *Wordnet. An electronic lexical database*, MIT Press, Cambridge, Mass.

GARDENT Claire, MANUELIAN Hélène, KOW Eric Y. (2003), Which bridges for bridging descriptions ? Actes de l'atelier *Linguistically Interpreted Corpora*, Association for Computational Linguistics, Budapest, Hungary.

GROSS Maurice (1975), *Méthodes en syntaxe* : Paris, Hermann.

GROSS Maurice (1991), Les banques de données du LADL : analyse automatique et couverture. *Actes du colloque "Informatique et langue naturelle"* p. 361-386, Nantes : LIANA.

ISO (2003)*, Lexical Markup Framework, proposition ISO TC37 / SC4* , accessible à la page : http://pauillac.inria.fr/atoll/RNIL/TC37SC4-docs/N089.pdf

MANGEOT – LEREBOURS M., SERASSET G., LAFOURCADE M., (2003), Construction collaborative d'une base multilingue. Le projet Papillon, Traitement Automatique des Langues, Volume 44, n°2/2003, Paris, Hermès.

MANUELIAN Hélène (2003), Descriptions définies et démonstratives : analyses de corpus pour la génération automatique de textes, Thèse de Doctorat, Université de Nancy2.

MANUELIAN Hélène, TIMMERMAN Carine (à paraître), Un projet du CNRS en cours de réalisation : L'informatisation du *Petit Larousse 1905* et d'une collection millésimée et séculaire.

MASMEJEAN Julie(2005) *Non-dits et mutisme socioculturels dans la langue française et dans ses dictionnaires du XIX au XXIème Siècle, autour de trois mots tabous : Erotisme, Pornographie, Sexualité*, Mémoire de Maîtrise, Université de Cergy Pontoise.

MUELLER M., *A very gentle introduction to TEI*, document Internet accessible à : http://www.tei-c.org/Sample_Manuals/mueller-main.

PRUVOST Jean (2004), *La dent-de-lion, la Semeuse et le* Petit Larousse, Paris, Larouuse.

SPERBERG-MCQUEEN, CM et BURNARD L (eds), 2004 Guidelines for Electronic Text Encoding and Interchange, XML-compatible edition : http://www.tei-c.org/P4X/index.html

## Remerciements

Merci à Jean PRUVOST pour son soutien dans le projet d'informatisation du *Petit Larousse Illustré* de 1905.

Un grand merci à Eric Y. KOW (LORIA) pour ses précieuses informations sur les technologies du Wiki, son aide autour de l'analyse syntaxique et du balisage des textes durant les quatre dernières années et sa relecture.