Pour une conception réticulaire des dictionnaires

Arguments pour une refondation de la structure des dictionnaires sur la base des nouvelles possibilités technologiques issues du *web* et des techniques associées

Yannis Delmas-Rigoutsos

ERT 49 – Ingénierie des Ressources Médiatiques pour l'Apprentissage
MSHS, Université de Poitiers, 99 avenue du Recteur Pineau, 86022 Poitiers cedex
IUFM de Poitou-Charentes, 22 rue de la tranchée, 86000 Poitiers
yannis.delmas@poitou-charentes.iufm.fr

RÉSUMÉ.
Cet article présente trois pistes de réflexion pour repenser les dictionnaires, encyclopédies et lexiques dans le cadre du web *sémantique en terme de micro- et de macro-structuration du corpus. Il définit et défend une notion d'accès direct aux articles qui le composent. Cet article s'appuie tant sur l'état de l'art des outils informatiques standards disponibles que sur les pratiques réelles de consultation du* web *(en France).*
Il comporte, par ailleurs en annexe une mesure de la référence de Wikipedia *dans* Google.

ABSTRACT.
The present paper introduce three ways to rethinking dictionaries, encyclopædiæ and lexicons in the semantic web framework. The author addresses both the micro- and the macro-structure of the corpus. *He defines and promote a concept of direct access (*accès direct*) to article constituting that* corpus. *He relies on the state of the art in computer science available standards as well as on the real use of the web (in France).*
Moreover, as an appendix, there is a measurement of the Wikipedia *reference in* Google.

MOTS-CLÉS : dictionnaire, dictionnairique, web, web semantique, XML, RDF, OWL.

KEYWORDS: dictionary, dictionarics, world-wide web, semantic web, XML, RDF, OWL.

Préalable terminologique

Dans cet article, nous entendrons par « dictionnaires », *stricto sensu*, les dictionnaires de langue et dictionnaires spécialisés, constitués d'articles visant à définir des mots ou expressions.

Toutefois notre propos dépasse ce seul emploi et vise, outre ceux-ci, également les thésaurus, encyclopédies, lexiques, dictionnaires étymologiques, dictionnaires historiques, de noms propres etc. Faute d'une expression plus appropriée, nous emploierons l'expression « dictionnaires », *lato sensu*, pour désigner ce genre littéraire. Sauf indication contraire, nous nous attacherons à ce second sens.

1. Introduction – la « révolution » d'Internet et du *web*

En quelques années seulement Internet et les technologies d'information et de communication (TIC) ont imposé à la société occidentale certains usages et méthodes de travail qui ont profondément renouvelé son rapport à l'information. Soulignons quelques uns des principaux aspects de cette « révolution » :

– Elle renouvelle le rapport au texte et fait éclater le cadre jusqu'alors quasi omniprésent de la linéarité, issue du *volumen* antique. La culture se vit désormais plus souvent « large » que « profonde », extensive qu'intensive. L'information, le savoir, sont « picorés » au gré de l'hypertextualité. La connaissance, aujourd'hui, doit être faite de **fragments courts**, clairement **identifiés** et abondamment **reliés** entre eux. Cette évolution du rapport aux textes d'information et d'apprentissage se double d'une accélération de notre rapport au temps de prise de connaissance. Ces deux évolutions sont amplifiées par les TIC mais débordent largement leur cadre, comme le montre la faveur considérable des ouvrages de culture scientifique, littéraire ou artistique en forme de dictionnaires.

– Elle renouvelle également les modes de relations humaines. En affichant des exemples de création commune éclatants par leurs qualités et par leur ampleur, en permettant le développement de logiciels et de savoirs libres et gratuits ou simplement *opensource*, Internet offre au grand public les standards d'organisation humaine de la communauté scientifique. Cette norme combine **confiance *a priori*** et possibilité de **contrôle *a posteriori***, d'une part, et diffusion généreuse du savoir et respect du droit d'auteur, d'autre part. GNU/Linux et Internet lui-même figurent parmi ses plus belles réalisations. Suivant cette tradition, nombre d'auteurs publient gratuitement par Internet quantité de logiciels et d'informations les plus diverses, de la recette de cuisine au cours d'astrophysique.

– Tout ceci ne pouvait qu'entraîner une modification profonde du rapport à l'expression écrite. Il n'est plus nécessaire aujourd'hui d'être « Écrivain » ou « Artiste », ni d'avoir le statut social correspondant, pour publier : les TIC autorisent désormais une production textuelle fondée sur des rapports de communication possiblement horizontaux. La notoriété d'une œuvre est

potentialisée par l'intérêt des lecteurs (pour de « bonnes » et de « mauvaises » raisons) alors que ce n'est plus le cas, dans une large mesure, de la publication papier, largement marchandisée et dont la propriété semble irrémédiablement concentrée. La publication électronique autorise, de ce fait, une prépondérance de la qualité de conception sur le jeu de rapports de force. C'est ainsi que les normes d'Internet ont pu se développer sur des critères de pertinence et de qualité technique et s'émanciper, dans une certaine mesure, de la défense d'intérêts particuliers (économiques ou politiques) qui parfois dominent les organismes de normalisation (ISO, AFNOR, ANSI…). Pour ces mêmes raisons, cette organisation offre également une large audience à la désinformation la plus fantasque.

Ce nouveau contexte technologique croise, depuis quelques années, un **besoin considérable de diffusion de la culture**, de l'information et des connaissances **scientifiques et techniques**, à tel point qu'on voit mal sur quelles bases le Citoyen francophone des démocraties modernes – pour n'évoquer que lui – peut éclairer son opinion concernant les grands choix politiques de son temps. Une traduction de ce fait est que les dictionnaires (*stricto sensu*) usuels n'atteignent pas le niveau du baccalauréat. Une manifestation plus grave, peut-être, pour le long-terme en est la désertion des filières scientifiques des universités et la grande difficulté, pour le moins, de recruter des enseignants dans certaines disciplines.

En réaction à ce besoin, nombre d'auteurs se sont saisis des moyens de publication du *web* pour la diffusion de connaissances. Certaines œuvres prennent actuellement la forme de dictionnaires (*lato sensu*) et en renouvellent ainsi le genre puisque, pour la plupart, leurs auteurs ne sont pas lexicographes.

L'auteur de cet article est responsable, dans un IUFM, de la formation aux TIC des enseignants du primaire et du secondaire et, par leur intermédiaire, de la formation de leurs élèves. À ce titre, et en tant que membre d'une équipe de recherche sur les usages des médias pour l'apprentissage, il a pu constater l'inadéquation entre les besoins manifestes et les moyens disponibles. Il porte, à ce titre, le projet d'un nouveau dictionnaire visant à combler les lacunes du domaine scientifique.

Cet article se place dans la perspective de ce projet mais ne s'y limite pas. Il veut prolonger la réflexion de nombreux auteurs de dictionnairique[1] qui ont souligné combien le nouveau contexte technologique autorise une **nouvelle conception des dictionnaires** et de l'accès à leur contenu et ont tenté de telles évolutions. Cet article plaide pour refonder la conception des dictionnaires dans le cadre d'un programme, appelé « ***web* sémantique** », qui vise à transformer le *web* en base de connaissances, au-delà de ce seul genre dictionnaire.

Nous aborderons d'abord le problème du mode réel d'accès à l'information contenue dans les dictionnaires, opposant deux types d'accès que nous nommerons « direct » et « par guichet ». Nous aborderons ensuite brièvement la macrostructure

[1]. Cf. p. ex.: [Zock & Carroll, 2003], [Fellbaum & Miller, 2003], [Hartrumpf & al., 2003], [Mangeot-L. & al., 2003], [Selva & al., 2003], [Mel'čuk & al., 1995, p. 52].

du dictionnaire et l'organisation de ses articles. Nous reviendrons ensuite sur la microstructure et le format actuel des définitions de dictionnaires et sur leur inadéquation au domaine scientifique. Enfin nous plaiderons pour l'enrichissement des articles par un réseau de liens répondant à un format standard du *web*, RDF. Nous conclurons en proposant une structure d'ensemble sur ces bases et en ouvrant quelques perspectives sur les champs d'action possibles insérant les dictionnaires en-ligne dans le cadre général de la constitution du *web* comme base de connaissances.

2. Guichet contre accès direct

2.1. *Accès par guichet*

Par soucis d'explicitation, nous commencerons par rappeler le procédé traditionnel d'accès à une ressource, que nous appelons « par guichet », afin de l'opposer, par la suite, à ce que nous appellerons l'accès direct.

Dans l'accès par guichet, la première étape pour consulter une **ressource** ou unité d'information, est de disposer d'une **référence**, ensemble de méta-données permettant de l'identifier de manière unique : type d'œuvre, titre, nom d'auteur, année de publication etc. Il faut ensuite accéder à un **fournisseur** d'accès (putatif) à la ressource en question : bibliothèque, médiathèque, librairie etc. Le fournisseur dispose d'un **guichet** où l'utilisateur demande l'œuvre, ou une copie intégrale ou partielle ([2]). Cette dernière est fournie à partir d'un **dépôt**, au moyen d'une cote : réserve locale, rayons, prêt entre bibliothèques etc.

Sauf dans quelques cas dégénérés[3], chaque étape est incontournable. Ainsi, il est impossible de disposer d'un ouvrage sans référence suffisante. Autre exemple : il est impossible de vérifier l'adéquation d'un ouvrage avant de l'avoir en main et donc de l'avoir requis auprès d'un fournisseur. Il est, bien entendu, nécessaire de savoir que tel guichet donne accès à tel ouvrage, sous peine de devoir reprendre la procédure *ab ovo*.

Cette procédure, découle de la forme matérielle des ouvrages et, à ce titre, s'est étendue des codex aux œuvres électroniques, cassette, CD, DVD etc., dans la mesure où celles-ci étaient d'abord distribuées sous forme matérielle. Désormais, alors que la publication immatérielle s'étend, par exemple pour les encyclopédies en ligne, ce mode d'accès reste quasi unique pour les publications d'éditeur. ([4])

[2]. Les bibliothèques à consultation en accès direct ne doivent pas dérouter le lecteur : leur utilisation constitue le plus souvent un accès « par guichet », dans notre terminologie.
Le guichet est implémenté par une recherche de localisation sur un plan des collections.
[3]. Au sens mathématique du terme. Par exemple, dans le cas d'une bibliothèque de fond de classe ne comportant qu'un nombre très limité d'ouvrages. Cas similaire : celui des usuels.
[4]. Notons que cela est également le cas pour les publications des fournisseurs d'accès à Internet. Ceci s'ancre probablement dans l'histoire de fournisseurs de services en ligne ou par téléphone de la plupart des plus grands d'entre eux.

2.2. *Accès par guichet – le cas des dictionnaires*

Dans le cas des dictionnaires, les unités d'information sont des articles. Le fournisseur est l'ouvrage dictionnaire. Le guichet est d'abord une recherche par ordre alphabétique de vedette. Puis, avec l'évolution de la technique, il se développe pour intégrer des outils informatiques répondant à des requêtes, souvent partielles et parfois inexactes. Ces outils permettent généralement une recherche plein-texte et peuvent être d'une grande complexité (pour les principaux ouvrages). On verra, par exemple, [Dendien & Pierrel, 2003, pp. 27 sqq.] pour le cas du *Trésor de la langue française informatisé*, TLFI, lequel permet, notamment, une recherche phonétique.

Il n'en reste pas moins que les dictionnaires d'éditeur restent, actuellement, dans le modèle de l'accès par guichet : pour chercher la définition d'un mot, ce qui reste l'usage essentiel d'un dictionnaire[5], il est nécessaire d'accéder d'abord à l'outil de recherche d'un dictionnaire particulier[6]. Dans le cas des domaines scientifiques il n'est pas rare qu'une définition ne convienne pas au lecteur ; il doit, dans ce cas, reprendre sa recherche *ab ovo* dans un autre ouvrage, à l'aide d'un autre guichet, et ainsi de suite jusqu'à obtenir satisfaction... ou abandonner.

2.3. *Organisation de l'information sur le* web

Afin de préciser (de contextualiser) ce que nous entendons par « accès direct », faisons retour sur l'histoire du *web* et de son modèle de publication. ([7])

Rappelons d'abord qu'**Internet** est un réseau d'interconnexion. Il agrège, sur la planète, quantité de réseaux plus petits, locaux ou régionaux. Chaque terminal connecté à Internet dispose d'une adresse numérique permettant de l'identifier parmi tous. Une telle adresse permet à un terminal d'héberger des services, potentiellement à l'usage d'Internet tout entier. Certaines adresses disposent d'un nom de domaine (DNS) permettant de les désigner de façon humainement lisible. C'est systématiquement le cas pour les serveurs de publication.

Chronologiquement, le premier service de publication est le transfert de fichier (**FTP**, 1970) à consultation anonyme[8]. Il disposait de dépôts appelés « serveur FTP anonymes ». Les ressources publiées sur ces serveurs étaient listées dans un ensemble de répertoires appelés « Archie » mentionnant le plus souvent seulement leur nom, leur chemin et leur date de publication.

---

[5]. Dendien et Pierrel [2003], sur la base de leur expérience du TLFI, constatent : « pour l'utilisateur “ grand public ”, la fonction essentielle d'un dictionnaire [de langue non spécialisé] est de vérifier le sens et l'orthographe d'un mot donné, la proportion d'utilisateurs procédant à des recherches transversales est infime ».

[6]. Il s'agit même parfois d'une recherche doublement par guichet puisque, quand il ne s'agit pas d'un usuel, le dictionnaire particulier lui-même doit être obtenu par guichet.

[7]. Pour plus de détails, cf. [Delmas-Rigoutsos, 2004, chap. 4], dont nous nous inspirons. Cf. [id., pp. 29 sq.] pour une bibliographie sur ces points. Pour une histoire des débuts d'Internet, cf. [Huitema, 1996].

[8]. La première version du FTP est publiée en 1970, mais d'autres systèmes similaires l'ont précédée. Cf. [Delmas-Rigoutsos, 2004, eod. loc. ; Huitema, 1996].

Le mot « hypertexte » est inventé en 1965 par Ted Nelson (projet *Xanadu*) mais reprend des idées remontant au moins à Vannevar Bush [1945] (concept de « memex »). Il est réellement mis en œuvre entre différents dépôts de documents par le système **Gopher** en 1991. Son système de répertoire, *Veronica*, a évolué pour être intégré au système d'hypernavigation de Gopher lui-même.

Tim Berners-Lee étendra le système à la même époque sous la forme du ***world-wide web***. Les dépôts, appelés serveurs HTTP, reposent, principalement, sur une simplification du FTP. Les liens sont désormais insérables à l'intérieur-même des documents publiés, dans HTML, dans PDF puis dans toute application de XML. Depuis l'implémentation de CGI/1.0 en 1994 les pages peuvent être produites dynamiquement, ce qui permet de disposer très tôt de recherches qui dépassent la consultation de listes fixes et ouvrent la voie vers la recherche multicritère sur le *web*. Les premiers répertoires thématiques, successeurs de Veronica, seront vite remplacés par des moteurs de recherche plein-texte.

Eu considéré « Les hyperliens, combinés à la recherche multicritère », pour Zock & Carroll [2003, p. 8], « [l]'informatique nous libère [...] de la camisole papier ». Certes, mais en transformant l'outil, et donc ses possibilités, ces techniques modifient également le comportement d'accès à l'information. Les étudiants actuels, de même que nombre d'utilisateurs chevronnés du *web*, accèdent à leur information le plus souvent à partir d'un moteur ou d'un « méta-moteur » de recherche. Cette information n'est pas toujours de la meilleure qualité, mais ce type de recherche autorise un certain nombre de justifications que nous nous proposons d'examiner.

2.4. *Accès direct aux ressources*

Un bon site *web* se doit d'être un site intriqué à la « toile » universelle. Les pages comportant peu d'hyperliens à valeur ajoutée[9] sont ainsi péjorativement reléguées au rang de « cul de sac », quel que soit leur intérêt documentaire propre. De fait, les moteurs de recherche accordent généralement une importance considérable à la valeur extrinsèque d'un document relativement à leur valeur intrinsèque.

Cette stratégie répond à deux nécessités de la recherche sur le *web*. *Primo*, l'immensité du *web* et la polysémie des mots font que le degré de pertinence effectif d'une réponse à l'utilisateur est faible et qu'il doit généralement parcourir les premiers items de celle-ci afin de déterminer lesquels lui conviennent. Une unité documentaire subit donc deux types de consultations : bien sûr la lecture, cursive ou rapide, mais, bien plus souvent, le survol d'évaluation de pertinence, souvent suivi de rebonds hypertextes. *Secundo*, pour évaluer précisément le champ sémantique d'une unité de petite taille, le moteur, comme le lecteur humain, doit rassembler ce qui est épars, doit agréger des informations puisées dans son hypercotexte.

9. Un aiguillage vers une partie de site (menu ou autre) n'a aucune valeur ajoutée documentaire en tant qu'hyperlien. C'est un simple jalon. Un lien de type « recherche plein-texte » ou « moteur de recherche », pour d'autres raisons, n'en a pas plus : il fournit simplement un raccourci pour une opération autrement possible. Les informaticiens nomment dans leur jargon « édulcorant » ce type de procédé, parfois utile néanmoins.

Dans le cas des dictionnaires d'éditeur la situation est pire encore que pour les textes généraux. Ceux-ci sont le plus souvent, en effet, en accès par guichet, selon notre terminologie, et, de ce fait, *c'est le dictionnaire en lui-même qui constitue une unité d'information*, plutôt que ses articles, qui, de fait, sont écrantés, invisibles aux robots des moteurs. Or, c'est bien « pour la compréhension d'un mot et donc pour la consultation des définitions que le dictionnaire monolingue est le plus utilisé »[10].

Appelons **accès direct** à l'information le fait d'accéder à des références d'unités d'information (du degré pertinent de granularité) avant toute considération de fournisseur et *a fortiori* de dépôt de données. L'accès aux « pages » *web* par moteur de recherche en est un exemple. Celui-ci montre qu'un tel accès peut dépendre d'un outil ; toutefois, celui-ci est généraliste, parfois même incorporé à l'outil de consultation (le navigateur), et ne constitue aucunement le guichet d'un fournisseur ou d'un dépôt. Les fournisseurs ne sont que rarement indiqués dans la réponse de ces moteurs de recherche et le dépôt, signalé par son URI, n'est bien souvent que l'une des (quelques) informations données au lecteur afin de lui permettre d'évaluer le degré de pertinence des réponses à sa demande. ([11])

Le moteur de recherche *web* n'est pas le seul cas d'accès direct. Les correcteurs orthographiques et grammaticaux de certaines applications bureautiques en sont également. Accès direct encore qu'une interface DICT[12] ou *Wikipedia*[13] dans une application bureautique, un bureau de système d'exploitation ou un bureau virtuel (BV) / environnement numérique de travail (ENT).

Le principal problème de l'accès par guichet, face à l'accès direct, dans le cas des dictionnaires est l'extrême dispersion de l'information. Pour que ce mode de recherche soit efficace, en effet, il faudrait se contenter de sources peu nombreuses et se passer de les confronter. Dans le domaine général de la langue, on peut effectivement se contenter du TLFI, par exemple. En revanche, si l'on souhaite des informations plus poussées, dans un domaine spécialisé, par exemple scientifique, les dictionnaires de langue sont insuffisants, les encyclopédies d'éditeur incomplètes, le *Grand dictionnaire terminologique* pas toujours fiable etc.

Outre ces manques, un dictionnaire, quel qu'il soit, ne peut qu'adopter un point de vue et ne donnera pas toujours une réponse pertinente pour une recherche fine

[10]. [Selva & al., 2003] d'après les observations de [Bogaards, 1988] et [Chi, 1998] (contexte d'apprentissage du français langue étrangère ou seconde).
[11]. Nous voulons, par cette terminologie, renvoyer à l'image des bibliothèques à accès direct, par opposition aux bibliothèques à guichet de consultation. Toutefois, cette métaphore ne doit pas être filée trop strictement.
[12]. Le protocole DICT [Faith & Martin, 1997] permet d'interroger des bases de données en forme de dictionnaire accessibles gratuitement sur Internet. Les principales bases disponibles aujourd'hui sont en anglais : *Webster's Revised Unabridged Dictionary* (1913), *WordNet* (anglais, v. 2.0, 2003), dictionnaires d'informatique et sur la bible, *Factbook* de la CIA (2002), etc. Les logiciels clients DICT sont généralement configurés pour faire une recherche simultanée dans plusieurs ouvrages et fournissent ainsi un accès direct. Pour la plupart, ils permettent également un accès par guichet, dictionnaire par dictionnaire.
[13]. La *Wikipedia* est une encyclopédie contributive libre sur le *web*. Cf. annexe 1.

sortant d'un simple cadre définitoire. Il reste essentiel de pouvoir confronter **différentes sources**. C'est d'ailleurs ce qui est systématiquement enseigné dans les cours sur la recherche d'information, de l'école primaire (B2I école) à l'Université (C2i niveau 1). Une recherche optimale, aujourd'hui, devrait donc combiner des ouvrages de référence d'éditeur, consultés par guichet, et une recherche d'information sur le *web* en accès direct, par moteur de recherche. Force est de constater que la plupart des utilisateurs réels négligent les ressources d'éditeur, malgré leurs qualités pédagogiques d'exposition et leur fiabilité, pour se contenter souvent de sources secondaires à l'autorité scientifique douteuse, pour le moins.

L'utilisateur non spécialiste de recherche d'information n'est pas capable de retenir une source pour l'orthographe élémentaire (*Orthonet*, p. ex.), une autre pour la langue courante (TLFI, p. ex.), une autre pour la terminologie spécialisée (GDT, p. ex.), une autre pour des articles encyclopédiques généraux (*Encyclopædia universalis*, p.ex.), une autre encore pour les abréviations etc.

Aujourd'hui certains de ces éléments, tels que la féminisation [Becker, 1999] ou la phonétisation [D'Alessandro & Tzoukermann, 2001], par exemple, pourraient être agrégés à tout ouvrage en-ligne, sans que celui-ci ait à être repensé et sans grande complication juridique. Mais tous ne le pourront pas. Il n'est que de penser aux ouvrages dont l'accès est payant et la structure de données non librement accessible. Les bases *WordNet*, par exemple, qui sont un outil de travail pour de nombreux linguistes, développées d'abord par et pour la recherche, sont d'accès payant[14]. Avec le développement des ENT scolaires, les initiatives ENÉE/ENS, CNS et KNÉ [Braun, 2004 ; SDTICE, s.d.] ont montré qu'une certaine agrégation de ressources d'éditeurs différents, voire en concurrence directe, est possible. Le développement de ce type d'outils est susceptible de contribuer à développer ces usages. On voit mal, toutefois, pourquoi un environnement scolaire intègre qui *Le Robert*, qui le *Larousse* mais pas le *Dictionnaire de l'Académie française* ni le TFLI ou le GDT, sauf à supposer, bien sûr, des missions... ou intérêts autres que strictement scolaires. Bien entendu, l'accès par guichet, contrairement à l'accès direct, s'oppose à la multiplicité des « usuels » disponibles, pour des raisons de charge mentale... ou de charge graphique. Techniquement, une véritable **intégration** (laquelle a montré son intérêt pédagogique [Dokter & al., 1998]), n'est possible que sur un accès direct ou, en cas d'accès par guichet, sur un nombre extrêmement limité de sources, idéalement une seule.

Pour conclure notre promotion de l'accès direct, nous insisteront sur une bonne pratique de conception de logiciels : la **modularité**. Les logiciels conçus de façon non modulaire sont généralement tout à la fois plus lourds, moins fiables, plus coûteux, moins évolutifs (donc moins pérennes) et intègrent plus difficilement des contributions latérales ou des dérivations. La modularité est une simple adaptation à la programmation, à l'algorithmique, des principes organisationnels de division du travail. Dit rapidement, la modularité requiert de séparer les problèmes séparables.

[14]. Cf., p.ex.: <http://www.elda.org/catalogue/en/text/M0015.html>.

Dans le cas des dictionnaires, on ne voit pas pourquoi les systèmes de recherche seraient attachés aux corpus. À titre d'exemple voir [Dendien & Pierrel, 2003], qui détaille un certain nombre de difficultés liées à la recherche dans le TLFI, problèmes d'orthographe, de diacritiques, de casse, de confusion phonétique *et cetera*, mais ne soumet aucun argument qui impose l'intégration ou association de l'outil de recherche à l'outil de dépôt, donc au corpus (ou à quelques corpus, dans ce cas). Toutes ces tâches peuvent être laissées aux moteurs de recherche généralistes, qui le font déjà fort bien et ont une pression considérable de leur clientèle pour le faire toujours mieux, et les lexicographes se concentrer sur le dictionnaire, son corpus et son mode de publication, à destination des humains et des moteurs. Ceci peut se faire d'autant mieux que les moteurs sont friands des dictionnaires en-ligne. Ainsi, quand une définition est présente dans *Wikipedia*, elle est souvent haut classée dans *Google*, malgré l'absence d'autorité scientifique de cette source[15].

La modularité peut être poussée plus loin encore et, dans le cas d'une recherche intra-corpus, séparer la désambiguïsation de contexte entre flexions, qui doit encore (largement) être assurée côté serveur, et l'utilisation de la portée et du contexte immédiat, qui, dans une certaine mesure, peut être assurée côté client.

Ajoutons enfin qu'après le développement des architectures logicielles trois-tiers, le développement des gros ensembles logiciels, en particulier les ENT mène à une **architecture** dite « **orientée services** » où le traitement est clairement séparé d'une couche de présentation des données. Les logiciels fournisseurs de services de traitement communiqueront avec un « présenteur » (*web*, WAP ou autre) dans des formats en cours de standardisation au sein du cadre XML : SOAP, XML-RPC ou autre XMLP [Fallside & Lafon, 2004]. ([16])

2.5. *Pour dépasser l'accès direct : l'accès guidé*

Umberto Eco [1998] présente, en forme d'humour, la télécommande comme l'invention majeure du xx^{e} siècle. Cause ou conséquence, cet objet est le symbole de

[15]. Nous démontrons ce fait en annexe : cf. annexe 1, ci-après.

[16]. L'évolution historique est la suivante : 1.– Logiciel monolithique intégrant le stockage et le traitement des données et l'interface utilisateur (ex. : traitement de texte, base de donnée personnelle MS Access...). 2.– Client/serveur : Un logiciel sur le serveur stocke et traite les données et gère la présentation des données, mais l'interface utilisateur est un logiciel, appelé « client » sur le terminal de l'utilisateur. Ce système permet le travail de plusieurs utilisateurs simultanément. 3.– Trois tiers : Un (ou plusieurs) serveur stocke les données, les traitements relèvent d'un logiciel intermédiaire (*middleware*) et l'affichage d'un client léger. Cette organisation permet d'optimiser et de standardiser le stockage de données. 4.– Architecture orientée services : des serveurs de données stockent les données, des serveurs de services reçoivent des ordres XML et sur cette base traitent les données, une couche de produit une interface et reçoit les demandes des utilisateurs, l'interface proprement dite est affichée sur un client léger sur un terminal de l'utilisateur. Cette architecture, plus modulaire, permet à plusieurs applications fonctionnelles (dans la couche de présentation) d'accéder aux mêmes données sans risquer de porter atteinte à des relations (fonctionnelles ou logiques) entre données (ex. : règles comptables ou réglementaires, réservation de train, etc.).

la volonté versatile de nos contemporains[17], du passage frénétique d'un présent rapidement non pertinent à un futur à évaluer. Pour reprendre une terminologie informatique, l'accès aléatoire à l'information tend désormais à l'emporter sur l'accès séquentiel, en particulier sur la lecture cursive[18]. Les usages de navigation sur le *web* montrent le même trait culturel, permis par une hypertextualité dense. De ce point de vue, le *web* est similaire aux ouvrages en forme de dictionnaire, en particulier aux dictionnaires proprement dits (*lato sensu*).

Bien entendu la structure logique du *web* favorise ces comportements, ou au moins les autorise. Il en va de même des dictionnaires qui, malgré une structure physique linéaire, ont une structure logique essentiellement réticulaire.

Pourtant, pour Pruvost [2000], « [le] problème technique posé par la place disponible dans un dictionnaire papier [...] a été lourd de conséquences, car il a d'emblée limité les dictionnaires à ne favoriser qu'un type d'accès au mot, l'accès formel, alphabétique. » Il regrette ainsi l'absence d'autres types d'accès, notamment onomasiologiques, analogiques etc. Et, comme le soulignent Selva & al. [2003] pour les dictionnaires électroniques, « [...] force est de constater qu'une fois parvenu à l'article, l'utilisateur retrouve la présentation traditionnelle des versions papier et nombre de leurs caractéristiques et inconvénients : grosse quantité de texte à lire pour les articles longs, enchevêtrement d'informations de différente nature [...], incohérence de la classification des sens [...], renvois hypertexte non ciblés, d'article à article au lieu de sens à sens [...] ».

Cependant, en 1945 déjà, Vannevar Bush [op. cit., sec. 8] entrevoyait, dans le cadre du « memex », le développement d'un nouveau métier : « *Wholly new forms of encyclopedias will appear, ready-made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified.* [...] *The chemist, struggling with the synthesis of an organic compound, has all the chemical literature before him in his laboratory, with trails following the analogies of compounds, and side trails to their physical and chemical behavior.* [...] *There is* ***a new profession of trail blazers****, those who find delight in the task of establishing useful trails through the enormous mass of the common record.* »

Au-delà des annuaires le besoin, la nécessité, d'un **accès guidé** s'est très tôt fait sentir sur le *web*. Comme en d'autres points de son histoire, il semble que la fonction engendre bien vite l'organe : la fin des années 1990 a été le théâtre d'une inflation considérable, démesurée, des pages uniquement constituées de renvois vers d'autres. Ces catalogues, frappés congénitalement d'obsolescence, montrèrent vite leurs limites et furent remplacés par des « sites portails », communautaires ou généralistes. Insistons que cet accès guidé n'était possible, en pratique, que grâce au système universel de cotes du *web* (les URI) associé au mécanisme de l'hyperlien de document à document, instrument de base de l'accès direct sur le *web*.

---

[17]. Nous n'évoquons ici, bien entendu, que nos contemporains *des pays riches*.
[18]. À n'entendre que qualitativement : nous n'avons connaissance d'aucune donnée quantitative. On gardera en mémoire que le principal *medium* reste encore la télévision.

3. Quel format pour les articles

3.1. *Mode de présentation de l'information*

Comme le souligne Pruvost [2000], les dictionnaires bénéficient d'une faveur considérable en France. Ainsi Quillet écrivait-il « La Bible d'aujourd'hui c'est le dictionnaire. À la Bible, les peuples anciens demandèrent la révélation ; au dictionnaire, les peuples modernes demandent la connaissance. » [cité in ibid., n. 1 p. 116]. Aujourd'hui, la situation a toutefois évolué et il faut parler davantage d'information que de connaissance : au-delà du dictionnaire *lato sensu*, nombre d'écrits sont seulement mis en forme de dictionnaire, c'est-à-dire composés d'un appareil d'articles courts rangés par ordre alphabétique de vedette. Ainsi en va-t-il, par exemple, d'un *Dictionnaire des sciences* (*op. laud.*), qui est d'abord un ouvrage culturel, d'un *Dictionnaire de cuisine*, qui est un manuel pratique, d'un *Dictionnaire du Nord-Pas-de-Calais*, qui est touristique, ou encore d'un *Dictionnaire des postures amoureuses*, pour un autre genre de transports... Au total, *Amazon.fr*, par exemple, fournit près de neuf mille références d'ouvrages dont le titre comporte le mot « dictionnaire », parmi lesquelles plus de la moitié est disponible[19].

À observer le reste du monde on s'aperçoit que cet engouement ne se cantonne pas à l'Hexagone. Il ne se cantonne pas non plus à la consultation. Nous en voulons pour preuve la progression des contributions à *Wikipedia* en langue anglaise : création en janvier 2001, plus de 20 000 articles un an après, 100 000 en janvier 2003, 200 000 en janvier 2004, 400 000 le 20 novembre 2004 ([20]). Et *Wikipedia* n'est qu'un exemple parmi d'autres, l'un des ouvrages-sources les plus volumineux, mais non le seul, loin de là. Y compris en français, le *web* regorge de ressources de type dictionnaire, lexiques, dictionnaires *stricto sensu* et encyclopédies, mais également de ressources définitoires et/ou pédagogiques sous d'autres formats, cours, polycopiés, exercices et exerciseurs. Enfin, on trouve de nombreux grands textes du domaine public publiés *in extenso*, notamment sous l'impulsion des groupements Gutenberg (tandis que les bases institutionnelles, type *Frantext*, sont, elles, d'accès restreint, donc par guichet, donc inaccessibles sauf à quelques initiés).

On le voit, la grande liberté du *web*, en particulier de format, n'a pas conduit à abandonner les genres traditionnels, selon nous non seulement pour des raisons de conservatisme. Par exemple, la longueur des articles n'augmente pas considérablement alors le *web* en offre la possibilité. On trouve même de nombreux lexiques spécialisés. La raison en est simple : à la contrainte d'espace s'est substituée une contrainte de temps, le temps de lecture admissible (cf. ci-dessus, §2.4). Elle s'applique *mutatis mutandis* aux autres cas pour lesquels la contrainte d'espace était traditionnellement invoquée, *du moins pour ce qui concerne le mode de présentation* ; en revanche, la libération de l'espace va permettre aux ressources *web* d'être moins sélectives que leurs équivalents papier.

[19]. Mesure effectuée par l'auteur le 21/11/2004 : précisément 8 714 et 4 437 titres resp.
[20]. Données de [Wikimedia, 2004] confirmées par notre estimation, cf. annexe 1 ci-après.

3.2. *Mode de structuration de l'information*

Structurellement, les dictionnaires d'éditeur ont d'abord été, et sont encore largement, des dictionnaires papier informatisés. Ceci explique le poids de l'histoire. Plusieurs auteurs, en particulier [Selva & al., 2003, op. cit.] ont proposé des pistes pour « rompre avec [cette] première génération de dictionnaires électroniques » en réfléchissant à ce que peut être un article de dictionnaire électronique, en particulier sur l'exemple du DAFLES[21]. Nous renvoyons à cet article pour ses diverses propositions d'améliorations, tout en retenant que nombre de celles-ci ne peuvent guère s'appliquer à des œuvres contributives dans la mesure où elles demandent de réelles compétences en linguistique ou en lexicographie pour leur mise en œuvre et ne peuvent généralement pas être dévolues à des logiciels. Ainsi, par exemple, la présentation systématique des schémas actanciels et des contraintes de sélection.

En revanche, d'autres peuvent s'appliquer très largement, par exemple l'option de deux définitions, une courte et une redondante, ou un « effort particulier [...] pour la différenciation des co-hyponymes et synonymes partiels ».

Au-delà de ces améliorations, il y a nécessité, également, de préserver l'acquis que représente la structuration textuelle des dictionnaires. Souvenons-nous que le principe du balisage du texte a été inventé par les lexicographes des XVII^e^ et XVIII^e^ siècles. C'est un comble, à l'heure où la balise devient, par les applications de SGML puis de XML, l'instrument par excellence du numérique en général et du *web* en particulier, de voir sa complète régression parmi les dictionnaires contributifs. Jusqu'à récemment ce balisage n'était guère aisé à réaliser au niveau logique. Aujourd'hui plusieurs outils d'aide à la saisie, notamment ceux utilisés pour les publications scientifiques[22], pourraient contribuer à remédier à ce problème. On peut également envisager une aide automatisée à la saisie.

3.3. *Segmentation des unités d'information – qu'est-ce que définir ?*

Du point de vue de l'utilisateur, une définition de dictionnaire *stricto sensu* n'est guère plus qu'une association d'un *definiendum* et d'un *definiens*. Le *definiendum* peut être qualifié (catégories grammaticale, prononciation etc.), de même que le *definiens* (domaine, ressources associées,...) et l'ensemble être complété par des corrélats, tout en restant essentiellement informatif. Une définition ne vise finalement bien souvent qu'à préciser le sens et l'usage d'un mot, ce qui est bien en deçà des attendus usuels du terme.

Nous ne reposerons pas au fond la question de ce qu'est une définition : des mers d'encre ont coulé sur le sujet depuis, au moins, l'époque scolastique et avec, au moins, le concours de domaines aussi variés que la philosophie du langage, la

[21]. *Dictionnaire d'apprentissage du français langue étrangère ou seconde*. Cf. [Cowie, 1999], [Leech & Nesi, 1999] et <http://www.kuleuven.ac.be/dafles>.
[22]. On verra, par exemple LODEL <lodel.org> et Cyberdoc (Université Lyon 2).

logique, la psycho-linguistique et l'épistémologie. On lira [Sabah, 1997] pour un point sur le sujet. On en revient toujours finalement à la question de ce qui constitue le sens d'un vocable, qui n'est pas tranchée, loin s'en faut. Pour notre propos nous nous contenterons d'aborder un aspect seulement de cette question en nous concentrant sur ses attendus opératifs, que nous aborderons en terme d'objectifs pragmatiques. Notre point-de-vue sera d'abord celui d'un logicien et notre perspective se restreindra à celle d'un dictionnaire du champ scientifique.

Le point de départ symbolique de la vision moderne de la notion de définition est probablement le programme de Hilbert. Posé à la fin du XIX^e^ s., son objectif était de mécaniser la production des mathématiques. Il a ainsi donné une perspective à des travaux préexistants et a conduit à l'axiomatisation des mathématiques. Même si le programme initial s'est avéré irréalisable, du fait du théorème d'incomplétude de Gödel et autres résultats apparentés, il a pu trouver un certain achèvement au travers de la théorie des modèles, issue notamment des travaux de Tarski [1933]. Cette théorie sémantique reprend l'idée de Frege [1892] d'une correspondance entre expressions et monde représenté. Tarski reviendra plusieurs fois par la suite sur le fait que cette théorie ne s'applique pas directement aux langues naturelles. Pour autant, nous y insistons, la langue scientifique n'est pas une expression naturelle, en particulier dans sa vocation à représenter le réel : c'est une langue artificielle dont les vocables sont forgés et leur signification arrêtée, même quand elle reste vague, faute de pouvoir être « axiomatisée », d'une façon ou d'une autre. Contrairement aux mots de la langue usuelle [Vygotsky, 1934], le sens des mots scientifiques ne se modifie pas en fonction de la situation même. ([23])

Dans ce cadre, la position d'ouvrage de référence d'un dictionnaire *stricto sensu* impose une approche conventionnaliste, en particulier en contexte de formation. L'encyclopédie et le lexique peuvent être plus descriptifs et/ou explicatifs. Or, les dictionnaires (*s.s.*) actuels visent plus à donner des précisions, à forger une image mentale au sens de Johnson-Laird [1983] qu'à fournir des critères de séparation. Bien sûr ce mode de rédaction fournit quand même une sémantique, sinon on voit mal pourquoi les dictionnaires (*s.s.*) continueraient d'être utilisés[24], mais celle-ci se révèle souvent insuffisante dans le cadre pédagogique (en sciences). Le manque le plus important concerne le lycée et le premier cycle universitaire, les niveaux plus élevés faisant, de toute façon, appel à d'autres types d'outils.

Cette insuffisance n'est pas seulement le fait d'un choix rédactionnel. Nous pensons qu'il peut également s'agir du mode de segmentation du dictionnaire en unités d'information. Sous sa forme traditionnelle, le dictionnaire est une

[23]. À reprendre les principales descriptions épistémologiques du savoir scientifique, le vocabulaire fait partie du noyau dur (*alias* matrice disciplinaire, *alias* paradigme) des théories scientifiques. Le sens de l'essentiel du lexique est fixé et intangible jusqu'à la prochaine révolution scientifique. Toutefois ce sens *ne peut pas* être expresse pour tout le vocabulaire. Bien qu'univoque, le sens de certains termes fondamentaux doit rester implicite.

[24]. On peut voir là un parallèle avec ce qui se produit en logique dans le cas, par exemple, de la logique des prédicats ou du λ-calcul, où la collection des usages définit intrinsèquement une sémantique, sur un mode qui rappelle celui des paires opposables en linguistique.

« description du lexique [...] sous la forme d'une énumération [...] de lexies[25] [...], dans laquelle chaque lexie est munie d'informations pertinentes » [Mel'čuk, 1995, p. 19]. Autrement dit, une entrée définit typiquement un mot. Or cela ne peut suffire. Nous avons déjà évoqué la question de l'enseignement des langues et de la différenciation des co-hyponymes et sens proches, qui impose de les rapprocher au sein d'un même article [Selva & al., op. cit.]. Nous évoquerons maintenant le cas des sciences où **il advient fréquemment que la définition d'un terme ne puisse être isolée de celle d'autres termes**. En mathématiques, par exemple, il est impossible de définir un monoïde sans parler de loi de composition associative. En chimie, quel serait l'intérêt de placer à un endroit le symbole « Na » et à un autre le mot « sodium » ? Plus simplement, pourquoi éloigner la jument du Cheval, l'audiomètre de l'acoumètre etc. ? Dans un autre domaine, pouvons-nous isoler « aigle de Meaux », « Bossuet » et « Bossuet, Jacques Bénigne » ?

Dit autrement, un dictionnaire du domaine scientifique ne peut pas raisonnablement être un dictionnaire de mots (*Wörter*). Ce doit être un dictionnaire de choses (*Sachen*). De toute façon, les moteurs de recherche rendent obsolètes les dictionnaires de mots dans la mesure où ils peuvent fabriquer un rendu fonctionnellement identique à partir d'un dictionnaire de choses. Pour cela, il faut et il suffit que chaque article soit associé à des vocables faisant office de **lemmes** (et chaque lemmes à un certain nombre de formes orthographiques). Ceci permet, de plus, de mettre en œuvre quelques finesses dont se passent volontiers les dictionnaires traditionnels mais qui peuvent avoir un intérêt pédagogique telles que, par exemple, la différenciation des synonymes (qui ne sont qu'exceptionnellement exacts) ou la distinction entre l'épicène d'espèce et ses instances[26].

4. L'intertextualité pour dépasser les définitions

Ces points concernant les articles nous semblent nécessaires ; ils ne sont pas suffisants : une définition est une caractérisation mais n'épuise pas une description de l'objet, sans parler des questions d'usage des mots. De fait, les dictionnaires traditionnels incluent d'autres éléments pertinents d'information. Nous voudrions ici insister sur le caractère réticulaire de certains de ces éléments.

4.1. *Introduction : des liens morphologiques, sémantiques ou d'usage*

Actuellement, plusieurs projets dictionnairiques intègrent des liens entre lexies. Ainsi les *Wordnet* [Fellbaum & Miller, 2003] des différentes langues intègrent-ils

[25]. Suivant la terminologie de Mel'čuk, une lexie (lexème ou phrasème) est un vocable (mot ou locution) « pris dans une seule acception et [muni] de tous les renseignements qui spécifient totalement son comportement dans un texte » [1995, pp. 56 sq.].

[26]. Par exemple le mot « Cheval » peut désigner aussi bien un Cheval mâle (un cheval) qu'un Cheval femelle (une jument). La langue courante ne fait guère la distinction, aussi porte-t-on rarement la majuscule au nom d'espèce. En revanche, la distinction peut être (ou n'être pas) utile en langue savante (biologie) ou technique (hippologie, hippisme).

des liens morphosémantiques : hyponymie, synonymie, antonymie, méronymie, troponymie et implication (*entailment*), l'un de ces liens (la synonymie) étant traité de manière plus élaborée afin de construire des classes d'équivalence, les « *synsets* » [Miller, 1994] [27]. Dans tous les cas, le principe est celui de nœuds, ici des lexies, reliés par des relations orientées (graphe sagital) et qualifiées (ou étiquetées). Ce principe existe depuis longtemps en philosophie du langage et psycho-linguistique sous le nom de « réseau sémantique ». De nombreux travaux de psycho-linguistique ont montré l'aspect structurant de ces réseaux, au niveau cognitif même[28]. De nombreux avatars de cette notion servent également en pédagogie.

D'autres types de liens sont développés par d'autres projets. Ainsi le DAFLES insiste-t-il sur les collocations et éléments associés [Selva & al., 2003] (ex., pour « doigt » : « se mordre les doigts », « montrer du doigt » etc.). Citons encore, dans la même veine le problème des lexies polylexicales, omniprésentes en science.

Revenons à notre projet de dictionnaire scientifique par l'exemple des clades, qui composent, en systématique moderne, la taxonomie du vivant. Les meilleures descriptions cladistiques fournissent au moins, pour chaque taxon : les clades ascendants (p. ex.: l'embranchement pour une classe, le genre pour une espèce), une description, une caractérisation par rapport aux clades frères en terme de traits distinctifs (p. ex.: les mammifères sont homéothermes) et des exemples de divisions. Les ascendants sont assez similaires à des hyperonymes, les divisions à des hyponymes et les clades frères à des co-hyponymes, mais les premiers ne se résolvent pas en les seconds. Plus, la taxonomie est, finalement, la confluence d'une organisation hiérarchique similaire à celle traditionnelle des catégories et d'une sémantique de traits similaires au système de Katz et Fodor [1963]. Elle ne s'éclaire vraiment que sous ces deux aspects. Les dictionnaires classiques (*s.s.*) se contentent de rendre la hiérarchie par des corrélats (explicites ou non) et des outils plus complets, comme *Wikipedia*, de matérialiser cette description par des hyperliens.

Les outils d'aujourd'hui, conceptuels et techniques, permettent de dépasser ce stade et de prendre en compte des réseaux sémantiques complets et quelconques : n'importe quelles deux lexies peuvent se voir mises en relation par une troisième et cette relation accessible tant à une consultation humaine qu'automatique, qui plus est dans un cadre unifié pour l'ensemble du *web* : le ***web* sémantique**.

4.2. *Le* web *sémantique*

Nous aborderons ce qu'est le *web* sémantique et son ambition par un détour. Pour se procurer un document, il est nécessaire d'en disposer d'une référence. Il s'agit souvent du titre et de l'auteur ou de l'auteur et de la date de publication. Parfois, pour opérer un choix parmi plusieurs ressources, il est utile de bénéficier

[27]. D'autres projets, p. ex. Multi-Net [Helbig, 2001], distinguent plus de types de relations, sur le même principe, ou ajoutent un étiquetage systématique par des « rôles cognitifs », cf. [Hartrumpf & al., 2003], permettant notamment de préciser l'usage ou de désambiguïser.
[28]. Pour une revue sommaire nous renvoyons à [Sabah, 1997]. Parmi les travaux plus récents, signalons l'importance acquise par le « *Landscape model* » de Van den Broek et al. [1999].

d'informations plus précises : niveau d'enseignement, prérequis pédagogiques, cible, taille des ressources, etc. Ces informations sur un ouvrage, plus généralement sur une unité d'information, sont appelées ses méta-données. Aujourd'hui, une unité présente sur le *web* sans méta-données adéquates sera souvent mal référencée, et une unité sans référencement ou mal référencée n'existera, pour ainsi dire, pas. Le projet de *web* sémantique vise à étendre (considérablement) la notion de méta-données et à associer à chaque unité d'information d'autres unités liées par une relation, ce lien étant qualifié par une relation décrite elle-même par une unité d'information. Une telle réalisation, si elle est suffisamment dense et riche, nous rapprochera du visionnaire « memex » de Vannevar Bush [1945, op. cit.]. Il s'agit en fin de compte de constituer vraiment le *web* (une partie du *web*) comme une façon de super-encyclopédie, ce qu'il n'est pas actuellement, contrairement à une image généreuse largement répandue. ([29])

Le principal objectif opérationnel de ce programme est de nous sortir de l'actuelle noyade informationnelle et, d'une certaine manière, de dépasser l'hyperlien « blanc », non qualifié. Les liens offerts aux lecteurs seront plus précis, donc auront de plus grandes chances d'être pertinents, ils donneront surtout bien plus d'information aux moteurs de recherche, leur permettant de construire une description plus élaborée de l'hypercotexte d'une page, donc d'en mieux cerner le contenu, le sens et donc la valeur (intrinsèque et extrinsèque). À l'idéal, l'objectif est que les moteurs puissent un jour proposer une recherche contextualisée.

Qui dit méta-données dit liste d'autorité, qui est appelée « terminologie » dans le domaine du *web* et est implémentée par un objet informatique appelé « ontologie »[30]. « [*We*] *have the technology available for realizing the Semantic Web, we know how to built terminologies and how to use metadata* » ; le travail le plus important, actuellement, est donc la construction de terminologies. Pour l'instant, ceci se fait domaine par domaine, mais doit ultimement converger.

Une autre perspective, parente, est de faire en sorte de pouvoir décrire des unités d'information réutilisables, par exemple dans un cadre scolaire, de façon à ce qu'elles puissent être agrégées dans des ensembles plus vastes (p. ex. des cours). Ceci nécessite de subdiviser des unités actuellement trop volumineuses (p. ex. manuel de physique de quatrième) en unité plus petites et de décrire très finement ces dernières. Les éditeurs de ressources éducatives numériques ont amorcé ce travail. Celui-ci n'est possible que parce que les données peuvent être décrites (et chaînées) dans un cadre standardisé de méta-données, cadre d'ailleurs en cours de refonte avec la fusion attendue sous peu des référentiels LOM et SCORM (courant 2005). Ce travail reste à conduire pour la galaxie des dictionnaires.

---

[29]. Pour suivre les travaux sur le *web* sémantique nous renvoyons au groupe de travail du W3C <http://www.w3c.org/2001/sw/Activity> et au site <http://www.semanticweb.org/>.
[30]. Cet outil est plus complexe qu'une simple terminologie, en fait, et permet la définition de traits contextuels fins. Pour une présentation, cf. <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html> et <http://www.semanticweb.org/knowmarkup.html>.

4.3. *Les nouveaux formats de structuration locale de l'information*

L'outil fondamental de construction d'un tel réseau sémantique trans-documentaire est le format de fichier XML [Bosak, 1997 ; Bosak & Bray, 1999], pour le rendu bien sûr, mais surtout pour la structuration de l'information. Premier point, XML intègre l'encodage universel Unicode (UCS), donnant accès à l'écriture de toute langue, de l'alphabet phonétique, des symboles mathématiques et techniques etc. Ensuite, les formats de documents, appelés « applications de XML » sont automatiquement extensibles[31]: ils peuvent contenir des formules au format MathML, des graphismes vectoriels au format SVG, des liens entre ressources au format RDF, des ontologies et nomenclatures au format OWL ou, plus généralement, n'importe quelle (autre) application de XML.

Les types de documents XML sont des « systèmes ouverts », par excellence, avec leurs avantages, mais aussi leurs inconvénients : en effet, rien ne limite l'invention des concepteurs de formats XML, aussi il est essentiel de se préserver d'une divergence complète des formats. Par exemple, l'extensibilité MathML, qui sert à transcrire les formules mathématiques, n'a d'intérêt que parce que ce format constitue un standard, ou au moins un consensus. Si chacun construit son format de représentation des mathématiques de son côté, les interprétations des données ne pourront que rester limitées à l'aire de tel ou tel format ou logiciel. La représentation dans un format XML n'a réellement de sens, outre l'effet de mode ou de « modernité », que si celui-ci utilise les structures de données standards.

Dans le contexte des dictionnaires, eu considéré les éléments précédents du présent article, le format XML le plus central est le RDF (*Resource Description Framework*), outil principal du *web* sémantique. Comme son nom l'indique, les balises RDF visent à décrire toute ressource présente sur le *web*, qu'il s'agisse d'un document XML ou d'un fragment de document. RDF permet, notamment, d'associer à toute unité d'information, des métadonnées, ou des catégories à l'intérieur d'ontologies. Il est, dans ce cas, associé au format de description d'ontologies appelé OWL, *Web Ontology Language*, lequel n'est qu'une spécialisation du RDF, comme l'est également le format RSS, *RDF Site Summary*, qui permet de décrire les fils d'information relevant de tout ou partie d'un site *web*. Ainsi, par exemple, si les clades qui composent une taxonomie des être vivants sont représentés par des « catégories » OWL (le terme consacré est « classe »), l'ascendance des taxons est représentée par l'inclusion des classe. Cette inclusion étant décrite par des balises RDF, elle peut être utilisée par des logiciels afin de tirer des inférences sémantiques. De la même façon, on pourra qualifier tel vocable d'inusité, péjoratif etc. simplement par l'appartenance du vocable à la notion /*inusité*/, /*péroratif*/ etc. pour peu qu'on l'aie constituée en classe OWL.

Les balises RDF permettent également de décrire une relation d'une unité d'information à une autre en étiquetant ce lien par une troisième. Dans un dictionnaire, cela peut servir à qualifier des liens morphologiques et sémantiques

[31]. Du moins, pour peu qu'ils ne bloquent pas explicitement cette possibilité.

riches, par exemple :

moulage	– action –>	mouler
meunier	– métier –>	moudre
mouchoir	– moyen –>	moucher
mouture	– résultat –>	moudre

Bien sûr, ces liens n'ont pas de raison de se limiter aux liens morphologiques, même si ce sont les plus fréquents. Ainsi :

patient	– acteur –>	souffrir
réginaborgien	– gentilé –>	Bourg-la-Reine

Autre cas particulier, la relation peut être identique à l'un de ses termes :

action	– action –>	agir
acteur	– acteur –>	agir
produit	– produit –>	produire

Avec une telle organisation des données, les unités d'information d'un dictionnaire se constituent d'emblée dans son réseau sémantique naturel. Toute relation qui vaut la peine d'être explicitée l'est simplement en RDF, ces données, disponibles pour les logiciels, étant simplement rendues visuellement dans les pages produites à l'intention des lecteurs humains.

Pour préciser encore notre propos, nous choisirons deux contre-exemples sur le format DML (*dictionary markup language*) développé pour le dictionnaire *Papillon*[32] [Mangeot-L. 2002], afin d'observer comment il aurait pu être placé dans le cadre standard du *web* sémantique[33].

Premier contre-exemple les « types de politesse », qui caractérisent les vocables sont explicitement décrits comme des valeurs (« neutre », « respectueux », « humilité », « poli ») d'un attribut « politenessType ». Comme ces valeurs sont décrites en dur dans l'un des schémas XML du format[34], il faudra réviser celui-ci si l'on souhaite un jour modifier ces valeurs, alors que la référence à un fichier OWL permet, d'une part, d'alléger le schéma XML et, d'autre part, de modifier celui-ci simplement pour toute évolution. L'externalisation des catégories en OWL est donc une mise en œuvre de la modularité. Si l'on se place du point-de-vue d'un logiciel, le problème de ces valeurs est qu'elles sont a-sémantiques : rien ne distingue informatiquement « poli » de « neutre ». En revanche, une description OWL permet de renvoyer, pour chaque catégorie, à une URI précisant sa signification.

Second contre-exemple, la catégorie grammaticale est catégorisée par un attribut « posType » valant « n.m. », « n.m.inv. » etc. Comme précédemment, et pour les mêmes raisons, il serait souhaitable de modulariser le format en faisant appel à une

[32]. Le projet *Papillon*, <http://www.papillon-dictionary.org>, est un dictionnaire multilingue faisant appel à la contribution bénévole (pour diminuer le coût de constitution de la base lexicale). Il intègre divers dictionnaires numériques, lesquels sont convertis en XML/UTF-8. Ce dictionnaire a été conçu pour s'adapter, notamment, au français et au japonais.
[33]. Nous observons ce format parce qu'il a l'avantage de respecter les standards scientifiques en étant publié *in extenso* sur le *web* : <http://www-clips.imag.fr/geta/services/dml>.
[34]. Cf. <http://www-clips.imag.fr/geta/services/dml/papillon_fra.xsd>.

ontologie OWL. Dans ce cas l'intérêt est également de pouvoir décrire chaque valeur en terme de traits grammaticaux, susceptibles d'être utilisé par un programme informatique, par exemple pour une aide à la saisie des différentes flexions du vocable. Si l'on étend le cadre à celui d'un dictionnaire scientifique, des valeurs telles que « nom vulgaire d'espèce » ou « nom latin de taxon » peuvent indiquer une demande de saisie d'ascendants et de divisions. De la même façon l'indication d'un toponyme peut prévoir une association RDF au gentilé. *Et cetera*.

Un dictionnaire complètement intégré aux standards du *web* sémantique doit permettre, finalement, la syndication de contenus par description RSS au même titre que les sites *web* d'information. Si les articles sont bien des unités d'information du *web* (en accès direct), une telle description RSS permettra également l'intégration de ces articles dans des ensemble dépassant les seuls dictionnaires, par exemple, le projet MuREN (mutualisation de ressources pour l'éducation nationale), qui devrait être lancé par le ministère chargé de l'éducation nationale en 2005 [35] et qui prévoit le référencement RSS de ressources d'intérêt pédagogique.

Ceci n'est qu'une première étape : il est, en effet, possible d'aller plus loin encore et de proposer des liens qualifiés vers des ressources d'autorité, quand celle-ci se structureront elles-même pour le *web* sémantique, par exemple des bases de données taxonomiques, chimiques, statistiques etc.

5. Conclusion : vers un dictionnaire *web* réticulaire

Reprenons, en guise de conclusion, les principaux points d'aboutissement du texte ci-dessus pour décrire un dictionnaire réticulaire idéal au vu de l'état de l'art.

Aujourd'hui l'accès direct s'impose pour les dictionnaires en-ligne, au moins pour ceux d'entre eux qui sont gratuits. Un tel accès confie aux moteurs de recherche généralistes l'accès aux différents articles. Ces articles deviennent ainsi des unités d'information du *web* à part entière. Le dictionnaire n'a plus alors vocation à la complétion. Il n'est plus contraint à l'uniformité des ressources et peut proposer, dans un même ensemble, aussi bien des définitions, des présentations historiques ou étymologiques, des tableaux statistiques, des explications encyclopédique, des informations d'usage etc. Pour la même raison, tout vocable d'une définition devient potentiellement la source d'un corrélat, *via* un moteur de recherche (cf. fig. 1). Un tel dictionnaire a vocation à être utilisé soit par le *web* soit *via* des protocoles spécialisés, DICT ou correcteurs orthographique d'environnement de travail, idéalement le traitement du corpus doit être réalisé dans une architecture orientée services.

Le corpus d'un dictionnaire moderne doit être structuré comme un dictionnaire de choses. C'est à l'interface utilisateur de fournir des informations de type

35. Communication orale de B. Sillard, directeur de la direction de la Technologie, au cours du salon *Educatec*, 19 nov. 2004.

« dictionnaire de mots » ou « dictionnaire de choses ». De ce fait et du fait des nécessités des explications scientifiques ce dictionnaire ne peut se contenter d'associer un *definiens* à un unique *definiendum* : les articles doivent pouvoir être des références pour plusieurs lexies, qui leur font office de lemmes. Ils devront également pouvoir confronter des sens proches ou opposés et des co-hyponymes ou, plus généralement s'adjoindre tout moyen utile à la compréhension du lecteur.

Figure 1. *exemple d'un lemme polylexical*

Enfin, pour participer pleinement au mouvement de structuration informationnelle du *web* appelé « *web* sémantique », le corpus d'un dictionnaire réticulaire se doit d'être composé dans une application de XML modulaire faisant appel aux standards reconnus internationalement : RDF, OWL, MathML etc. Les articles doivent être riches en liens. Ces liens seront des hyperliens usuels mais également des liens RDF : morphologiques, sémantiques ou spécialisés. Les liens de chaque article doivent refléter le réseau sémantique existant entre ses lemmes. Chaque article doit pouvoir disposer de métadonnées décrivant précisément son apport. Le *web* sémantique visant à constituer une véritable encyclopédie trans-documentaire sur le *web*, tous les éléments qui constituent le dictionnaire, et non seulement son corpus, doivent adhérer à son principe de description. Pratiquement, pratiquement tout trait ou caractère des lexies ou des articles devra être un lemme du dictionnaire lui-même : types de liens, catégories grammaticales, domaines scientifiques, qualités des vocable etc. Plus encore, les références bibliographiques ou les contributeurs doivent pouvoir être décrits par le même procédé.

Pruvost [2000, op. cit., p. 24] nous dit que « [...] le fait d'être passé du lecteur intensif, celui qui retient tout, au lecteur extensif, celui qui consulte tout, pour accéder [...] au lecteur actif, celui qui peut marquer son empreinte sur le texte [...],

auxquels s'ajoute désormais le lecteur “ planétaire ” [...] n'est pas toujours synonyme de progression systématique ». C'est que le domaine a besoin d'être urbanisé, de façon résolue, par des experts des dictionnaires, des experts des domaines concernés et des experts des TIC, notamment. Nous ne pouvons nous contenter d'auto-organisation, sauf à nous exposer au risque de la régression.

C'est dans l'espoir d'offrir une perspective d'organisation que nous avons proposés ces quelques éléments de réflexion. Bien entendu, nous ne prétendons pas couvrir tous les éléments que devrait incorporer un dictionnaire réticulaire bénéficiant de tout l'état de l'art à ce jour. Il y aurait ainsi beaucoup à dire, par exemple, sur l'intégration à l'*Open archives initiative* (OAI)[36], les voies de collaboration avec les éditeurs des moteurs de recherche afin de proposer des canaux de données pertinents ou inversement d'extraire des données de l'immense corpus dont ils disposent ou sur les relations avec les recherches en intelligence artificielle[37].

6. Annexe 1 : analyse du référencement dans *Google* des données de *Wikipedia*

6.1. *Contexte*

Wikipedia est un dictionnaire encyclopédique contributif sur le *web*. Les contributions ne sont pas validées. Nous observerons ici sa version anglophone <http://en.wikipedia.org>. Ce projet est précédé par le projet *Nupedia* (début en mars 2000), similaire mais à comité de lecture. *Wikipedia* débute en janvier 2001, face au manque de contributions à *Nupedia*. L'historique de *Wikipedia* dans *Wikipedia* signale le passage du cap des 300 000 articles début juillet 2004. Nous estimons actuellement ce nombre à plus de 400 000 (cf. infra).

L'objet de cette annexe est d'estimer la position des articles de *Wikipedia* dans les références *Google*. Nous formulons l'hypothèse d'une position excellente.

6.2. *Protocole expérimental*

Toutes les données brutes ainsi que les logiciels de traitement (pour chaque étape) seront disponibles à partir du site *web* de l'auteur.

1.– *Wikipedia* donnant un accès libre à toutes ses ressources, nous l'avons explorée à la recherche de toutes ses pages. Le 16/11/2004 nous avons compté **708824 entrées brutes**, hors noms spéciaux qui ont été écartés[38].

[36]. Cf. <http://www.openarchives.org>.
[37]. Par exemple pour établir des liens RDF décrivant les prototypes ou la typicalité [Rosch, 1975], ou encore pour structurer les informations de production syntaxique ou sociale [Bernicot, 1992].
[38]. Pratiquement, on été écartés : « Main_Page », « Current_events » et tout URL comportant le caractère « : », ce qui a pu amener à négliger quelques pages admissibles.

2.– Ont ensuite été écartées semi-manuellement les pages « administratives » de *Wikipedia*, en particulier toutes celles comportant ce nom ou un dérivé. Restent **708335 entrées propres**.

3.– À raison d'une seconde par entrée, nous aurions atteint une durée de huit jours de téléchargement si nous avions voulu les traiter toutes. Nous avons donc choisi de conduire une étude sur un **échantillon tiré au hasard** (fonction « rand() » de PHP 5). N'ont été conservées que 1 000 entrées, sans redondance. Aucun diacritique ni chiffre n'étaient admis dans l'écriture de la vedette. Ce critère a conduit à rejeter 136 entrées sur 1 136, soit 12 %.

4.– Nous avons ensuite récolté diverses informations, par téléchargement de toutes les pages de l'échantillon depuis *Wikipedia*, en particulier si l'entrée est une redirection vers un article sous une autre vedette.

5.– Les entrées qui n'étaient pas des redirections ont été cherchées dans Google jusqu'à concurrence du rang 100. De nombreux articles de *Wikipedia* (tous ?) sont repris intégralement par *theFreeDitionary.com*, nous avons donc également collectées les références à ce dictionnaire (lequel est généralement bien référencé). Les pages susceptibles de désigner une localité sont pré-repérées automatiquement (cf. étape 7)[39].

6.– Dans la mesure où ce dictionnaire reprend également d'autres sources, dont Wordnet 2.0, toutes ces références à *theFreeDitionary.com* ont été corrigées et téléchargées afin de constater si elles sont tirées ou non de *Wikipedia*. Ce constat est dressé sur une mention explicite par *theFreeDitionary.com*.

7.– *Wikipedia* intègre un certain nombre de descriptions de localités nord-américaine issues automatiquement de bases de données. Dans la mesure où, *primo*, seules les données originales nous intéressent et où, *secundo*, *Google* de doit généralement pas placer ces pages en bonne position, nous marquons manuellement toutes les pages désignant ces localités (cf. étape 5). Nous isolons également les abréviations, pour lesquelles nous ne pouvons justifier le présent protocole.

8.– Traitement statistique du rang dans *Google* de ces données.

6.3. *Résultats*

6.3.1. *Nombre d'entrées*

La répartition entre catégories est la suivante : 317 redirections, soit 32 %, 83 localités nord-américaines, soit 8 %, 8 abréviations, soit 1 %, restent 592 pages normales présentant des définitions (dans un sens très large) ou des listes hors abréviations, soit 59 %.

On rappelle que le nombre d'**entrées brutes** utilisateurs était, à la date du recensement, de **708335**. On peut donc estimer, par extrapolation de l'échantillon,

[39]. Le traitement automatique se base sur la présence de références à <www.city-data.com>, <www.uscitydirectories.com> ou <www.stateguide.com>. Une vérification (étape 7) montre que ceci est suffisant pour les villes étatsuniennes. Nous n'aurons à ajouter manuellement (étape 7) que les cantons (*counties*) étatsuniens et les localités canadiennes.

un **nombre d'articles**, hors redirections de **près d'un demi-million** (extrapolation : 483 793) dont plus de 400 000 de définition ou liste (extrapolation : 425 001).

6.3.2. *Dynamisme des mises à jour*

Le tableau 1, ci-après, donne la répartition par déciles des dates de dernière modification des pages de la *Wikipedia* examinées au 20/11/2004. La dernière modification date du jour-même (D_{10}). La médiane (D_5) se situe à peine un mois avant. Autrement dit, plus de 50% des pages ont été modifiées durant le dernier mois, et plus de 20% durant les dix derniers jours (D_8). ([40])

décile	valeur
0	25/02/02
1	30/12/03
2	27/05/04
3	13/07/04
4	26/08/04
5	25/09/04

décile	valeur
5	*25/09/04*
6	17/10/04
7	02/11/04
8	11/11/04
9	15/11/04
10	*20/11/04*

Tableau 1. *Répartition des entrées par date de dernière modification*

6.3.3. *Références dans Google*

rang	signification	nombre	proportion des pp. référencées*	proportion des pages
1	première réponse	87	20 %, soit 1/5	15 %, soit env. 1/7
≤ 3	réponse 1 à 3	134	30 %, près de 1/3	23 %, soit env. 1/4
≤ 10	première page	236	53 %, plus de ½	40 %, soit 2/5
≤ 20	1re ou 2e page	304	69 %, plus de 2/3	51 %, soit ½
≤ 100	pages 1 à 10	443	100 % (par définition)	75%, soit ¾
> 100	non-référencée*	149	n/a	25%, soit ¼

Tableau 2. *Référence aux données Wikipedia dans Google*

[40]. Ces résultats correspondent aux données totale. Les résultats pour la seule catégorie « normale » sont encore plus massifs : D_5 = 28/10/04, D_8 = 12/11/04, D_{10} = 20/11/04.

Le tableau 2, ci-avant, donne le nombre de pages « normales » de *Wikipedia* (ou de *theFreeDictionay.com* reprises de *Wikipedia*) pour différents rangs d'apparition dans *Google* (au sein de l'échantillon décrit ci-dessus). On constate que deux pages sur cinq sont référencées[41]* dès la première page de réponses. Cela correspond à plus de la moitié des pages référencées*. Près du tiers des pages référencées* apparaît au rang 3 ou avant. Enfin, plus d'une page sur sept apparaît en première position (soit plus d'un tiers de celles qui apparaissent en première page).

7. Bibliographie ([42])([43])

Becker A. & al, *Femme, j'écris ton nom… Guide d'aide à la féminisation des noms de métier, titres, grades et fonctions*, Paris, La documentation Française, 1999.
NOTE. – Outil en ligne : <http://www.atilf.fr/feminisation> (s.d.: 2004/11/14)

Bernicot J., *Les actes de langage chez l'enfant*, Paris, PUF, 1992.

Bogaards P., « À propos de l'usage du dictionnaire de langue étrangère », *Cahiers de lexicologie*, vol. 52, 1988, pp. 131-152. Cité *in* [Selva & al. 2003] (r.s.).

Bosak J., « XML, Java, and the future of the Web », Sun Microsystems, 1997.
<http://sunsite.unc.edu/pub/sun-info/standards/xml/why/xmlapps.html > (1997/03/10)

Bosak J., Bray T., « XML and the Second-Generation Web », *Scientific American*, mai 1999. Paru en français sous le titre « Le langage XML », *Pour la science*, vol. 261, juillet 1999, consultable sur <http://www.pourlascience.com>.

Braun G., « L'offre numérique : CNS et KNE », *Educnet*, Ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche, direction de la technologie, sous-direction des TICE, avril 2004.
<http://www.educnet.education.fr/res/ens.htm> (2004/09/03)
CNS : <http://www.cns-edu.net/> (s.d.: 2004/11/20)
KNÉ : <http://www.kiosque-edu.com/> (s.d.: 2004/11/20)

Bush V., « As We May Think », *The Atlantic Monthly*, juillet 1945.
<http://www.csi.uottawa.ca/~dduchier/misc/vbush/awmt.html> (1994/04/27)
NOTE. – Sur « memex » cf. sec. 6 sqq.

Chi M. L. A., « Teaching dictionary skills in the classroom », *Proceedings of the 8th International Euralex Congress*, 1998, pp. 565-577. Cité *in* [Selva & al. 2003] (r.s.).

Cowie A. P., *English Dictionaries for Foreign Learners. A history*, Oxford, Clarendon press, 1999. (r.s.)

[41]. Par convention nous dirons « référencé » pour « référencé avec un rang au plus 100 » (cf. étape 5 du protocole expérimental). Nous rappellerons ce fait par une astérisque.
[42]. Toute URL est accompagnée d'une date : dernière modification du fichier quand elle est connue, sinon date de notre dernière consultation précédée de l'indication « s.d. » (sans date).
[43]. Les références correspondant à des renvois issus de citations et non matériellement utilisés par l'auteur sont indiqués « r.s. » (référence seconde).

D'Alessandro Chr., Tzoukermann É. (éd.), *Synthèse de la parole à partir du texte*, n° spécial de *Traitement automatique des langues*, vol. 42, n°1, 2001.

Delmas-Rigoutsos Y., *Internet et ses services*, support de cours du Master Webmestre éditorial, Université de Poitiers, 2004. (2004/10/05)
<http://yannis.delmas-rigoutsos.nom.fr/documents/YDelmas-services_Internet/>
<http://yannis.delmas-rigoutsos.nom.fr/documents/YDelmas-services_Internet.pdf>

Dendien J., Pierrel J.-M., « Le Trésor de la Langue Française informatisé », *Traitement automatique des langues*, vol. 44, n°2, 2003, pp. 11-37.

Dokter D. A., Nerbonne J., Schurcks-Grozeva L., Smit P., « Glosser-Rug : a user study », *in* Jager S., Nerbonne J., van Essen A. (éd.), *Language Teaching and Language Technology*, Lisse, Swets & Zeitlinger, 1998, pp. 167-176.
<http://odur.let.rug.nl/~glosser/Publications/userstudy.ps> (1997/07/10)

Eco U., *Comment voyager avec un saumon*, Grasset, 1998.

Faith R., Martin B., « A Dictionary Server Protocol », *Internet RFC/ STD/ FYI/ BCP*, RFC 2229, IETF, The Internet Society, 1997.
<http://www.faqs.org/rfcs/rfc2229.html> (2004/08/01)

Fallside D., Lafon Y., « XML Protocol Working Group », groupe de travail, *World-wide web consortium*, 2004. <http://www.w3.org/2000/xp/Group/> (2004/11/17)

Fellbaum Chr., Miller G. A., « Morphosemantic Links in WordNet », *Traitement automatique des langues*, vol. 44, n°2, 2003, pp. 69-80.

Frege G., « Über Sinn und Bedeutung », *Zeitschrift für Philosophie und Philosophie-Kritik*, vol. 100, 1892, pp. 25-50.

Guillot M.-N., Kenning M.-M., « Electronic Monolingual Dictionaries as Langage Learning : a Case Study », *Computers Education*, vol. 23, n°1/2, p. 63-73. (r.s.)

Huitema Chr., *Et Dieu créa l'Internet*, Paris, Eyrolles, 1996.

Hartrumpf S., Helbig H., Osswald R., « The Semantically Based Computer Lexicon HaGenLex », *Traitement automatique des langues*, vol. 44, n°2, 2003, pp. 69-80.

Johnson-Laird P., *Mental models*, Cambridge University Press, 1983.

Katz J., Fodor J., « Structure of a semantic theory », *Language*, vol. 39, 1983, pp. 170-210.

Leech G., Nesi H., « Moving towards perfection : the learner's (electronic) dictionary of the future », in : Herbst T., Popp K. (éds.) *The Perfect Learner's Dictionary (?)*, Tübingen, Max Niemeyer Verlag, pp. 295-306. (r.s.)

Mangeot-Lerebours M., « An XML Markup Language Framework for Lexical Databases Environments : the Dictionary Markup Language », *LREC Workshop on International Standards of Terminology and Language ressources Management*, Las Palmas (Canaries), mai 2002, pp. 37-44.
NOTE. – Pour tout détail, consulter directement le schéma XML du DML :
<http://www-clips.imag.fr/geta/services/dml> (s.d.: 2004/11/11)

Mangeot-Lerebours M., Sérasset G., Lafourcade M., « Construction collaborative d'une base lexicale multiligue – Le projet Papillon », *Traitement automatique des langues*, vol. 44, n°2, 2003, pp. 151-176.

Mel'čuk I., *Cours de morphologie générale (théorique et descriptive). Vol. 1. Introduction et première partie : Le mot*, Montréal & Paris, Presses de l'Université de Montréal & CNRS, 1993, 412 pp. (r.s.)
NOTE. – théorie « sens-texte ».

Mel'čuk I., Clas A., Polguere A., *Introduction à la lexicologie explicative et combinatoire*, Louvain-la-neuve, Duculot, 1995.

Miller G. A., Beckwith R., Fellbaum C., Gross D., Miller K., « Five Papers on WordNet », Cognitive Science Laboratory, Princeton Un., 1993.
<ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps> (1994/05/17)
<ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.pdf> (1997/07/31)
URL originale : <ftp://clarity.princeton.edu/pub/wordnet/5paper.ps> (plus disponible).

Pruvost J., *Dictionnaires et nouvelles technologies*, PUF, 2000.

Rosch E., « Cognitive representations of semantic categories », *Journal of experimental psychology : general*, vol. 104, 1975, pp. 192-233.

Sabah G., « Le sens dans les traitements automatiques des langues », *Traitement automatique des langues*, vol. 38, n° 2, 1997, pp. 91-133.

SDTICE, « Espace numérique des savoirs », *Educnet*, Ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche, direction de la technologie, sous-direction des TICE, s.d.
<http://www.educnet.education.fr/ENS/> (2003/04/28)

Selva Th., Verlinde S., Binon J., « Vers une deuxième génération de dictionnaires électroniques », *Traitement automatique des langues*, vol. 44, n°2, 2003, pp. 177-197.
<http://www.kuleuven.ac.be/grelep/publicat/tal.pdf> (2004/04/01)

Tarski A., « Pojęcie prawdy w językch nauk dedukcyjnych », *Wydział III Nauk Matematyczno-fizycznych*, prace Towrzystwa Naukowego Warszawskiego, vol. 34.
NOTE. – Article historique, traduit en allemand (avec *addendum*) en 1935, puis repris en anglais en 1944, sous une forme qui subira de nombreuses réimpressions et traductions.

Van den Broek P., Young M., Tzeng Y. & Linderholm T. « The landscape model of reading: Inferences and the online construction of a memory representation », in Van Oostendorp H. & Goldman S. R. (dir.), *The construction of mental representations during reading*, Lawrence Erlbaum Associates, Mahwah (USA, NJ),1999, pp. 71-98.

Vygotsky L., *Pensée et langage*, Paris, Éditions sociales, 1934.

Wikimedia foundation, « History of *Wikipedia* », *Wikipedia*, Wikimedia foundation, 2004
<http://en.wikipedia.org/wiki/History_of_Wikipedia> (2004/11/20)

Zock M., Carroll J., « Les dictionnaires électroniques », *Traitement automatique des langues*, vol. 44, n°2, 2003, pp. 7-10.
<http://telechargement.lavoisier.fr/TAL44_2_001-Zock_intro.pdf> (2004/10/19)