À la découverte d'un trésor : le mariage de l'informatique et de la lexicographie au service de la valorisation de la langue française.

Jean-Marie Pierrel

À la découverte d'un trésor : le mariage de l'informatique et de la lexicographie au service de la valorisation de la langue française

Jean-Marie Pierrel

1. Introduction

Depuis le 16e siècle, chaque siècle a donné naissance à au moins un grand dictionnaire de référence, dictionnaire de langue ou dictionnaire encyclopédique, reflet de la langue de l'époque. Ainsi pour le français on peut citer, entre autres : le *Dictionarium latinogallicum* de Robert Estienne (1552), le *Thresor de la langue françoyse* de Jean Nicot (1606), le dictionnaire dit de Trévoux (18e siècle), le *dictionnaire historique et critique* de Pierre Bayle (1740), le *dictionnaire critique de la langue française* de Féraud (1789), le *Littré* 1863-1872), le *Trésor de la Langue Française* (TLF) sans oublier le dictionnaire de l'Académie française et ses diverses éditions dont les plus remarquables, les 1re (1694), 4e (1762), 5e (1798), 6e (1835), 8e (1932-1935) et 9e (1992-...) éditions.

Dans ce domaine, la Lorraine a su prendre une place de choix comme en témoigne, par exemple, la publication en 1740 par le libraire Antoine d'une édition remarquable du Trévoux « dédiée au Roy de Pologne, Duc de Lorraine et de Bar », Stanislas, fondateur de notre noble assemblée.

Au cours de la seconde moitié du 20e siècle, de nombreuses contributions à la lexicographie française se sont développées à Nancy. Initiées au départ autour du projet de *Trésor de la Langue française* (TLF : Imbs, Quamada 1971-1994), les études nancéiennes en lexicographie française se sont poursuivies, au-delà de la rédaction du TLF, suivant deux orientations complémentaires (Pierrel & Buchi 2009) : la lexicographie historique et la valorisation informatique des ressources lexicales avec, entre autres, la valorisation d'une version informatique du TLF (Atilf 2004-2005), aujourd'hui disponible sur le Web (www.atilf.fr/tlfi), et le portail lexical du CNRTL (Centre National de Ressources Lexicales et Textuelles : www.cnrtl.fr) mis en place au sein de l'ATILF.

Cette seconde orientation a provoqué sur le plan des études lexicales une véritable révolution qui fit de l'informatique un outil indispensable pour :

- étudier le lexique et ses propriétés à travers l'exploitation intelligente de textes et de documents ;
- structurer et normaliser les connaissances lexicales et lexicographiques ;
- valoriser, partager et mutualiser les résultats de la recherche sur le lexique de notre langue, trop souvent encore dispersés.

Le partage et la mutualisation de résultats de recherche et de ressources informatisées sur le lexique français ouvrent en effet des perspectives intéressantes. La version informatique du TLF, sous forme de Cédérom et de ressources librement accessibles sur le Web, a rencontré un succès important tant auprès du grand public que des utilisateurs universitaires ou des professionnels de la langue. Référencé par d'innombrables sources, le TLFi fait l'objet de plusieurs centaines de milliers de connexions quotidiennes en provenance de tous les continents, devenant ainsi un dictionnaire incontournable et un outil de promotion appréciable de la langue française.

L'intégration plus récente au sein du portail lexical du CNRTL (www.cnrtl.fr/portail) de diverses ressources sur le lexique français permet encore une meilleure mutualisation des résultats de la recherche et une mise à disposition de nos résultats de recherche auprès de l'ensemble de la société. Aujourd'hui le portail lexical du CNRTL fait l'objet de plus de 300 000 requêtes par jour provenant d'horizons très divers (cf. : www.cnrtl.fr/aide/stat/): c'est l'un des sites Web sur le lexique français les plus utilisés.

2. Le TLF : une des dernières grandes aventures lexicographiques

Le TLF, *Trésor de la Langue Française*, est le fruit d'une des dernières grandes aventures lexicographiques qui, sous les directions successives du Recteur Paul Imbs et du Professeur Bernard Quémada, regroupa à Nancy plus de cent collaborateurs durant 30 années au sein du Centre de Recherche pour un Trésor de la Langue Française (CRTLF) puis de l'Institut National de la Langue Française (INaLF) dont notre laboratoire ATILF (Analyse et Traitement Informatique de la Langue Française[1]) se veut être aujourd'hui le digne successeur.

C'est en effet par décision du 20 décembre 1960 qu'est créé à Nancy un centre de recherche avec mission de mettre en œuvre la documentation, la rédaction et la publication d'un Trésor de la Langue Française. Ce choix avait été préparé et souhaité trois ans plus tôt dans les conclusions d'un colloque international de lexicologie et lexicographie françaises et romanes, organisé au Centre de Philologie Romane de Strasbourg par son directeur le Professeur Paul Imbs.

A cette époque le *Littré* (1863-1873) était tombé dans le domaine public et une intéressante controverse démontra que s'il était sage de réimprimer ce dictionnaire tel quel en raison de son caractère de « monument » de la science de son temps, le moment était venu de mettre en chantier quelque chose d'entièrement nouveau, qui tiendrait compte des acquis de la lexicologie et de la lexicographie du 20e siècle, des possibilités nouvelles en matière de documentation, et bien sûr des changements survenus dans la langue française depuis le milieu du 19e siècle. Les conclusions du colloque de Strasbourg étaient très nettes sur ce point : « *Instrument de travail, le Trésor poursuivrait donc un double but : être le témoin objectif et impartial du vocabulaire français, mieux connu parce que mieux inventorié ; être ce qu'avait été le Littré pour son temps : un exemple-type de lexicographie scientifique moderne* ».

Si l'on tente aujourd'hui de resituer le *Trésor* par rapport aux exigences de la lexicographie, rien de tel que de reprendre les objectifs initiaux définis par Paul Imbs. Le *Trésor* sera donc :

- Un *dictionnaire du monde francophone*. La France avait en effet sur ce point à rattraper un retard, à un moment où l'Angleterre avait terminé, vingt-cinq ans plus tôt, son New English Dictionary (Dictionnaire d'Oxford) et où plusieurs pays, latins, germaniques ou slaves, étaient à l'œuvre depuis plusieurs années pour publier un dictionnaire national.
- Un *dictionnaire historique*. Le Trésor ne se bornera pas à donner pour les mots l'usage du moment mais il inclura, pour chaque mot, une rubrique « étymologie et histoire », riche des connaissances actuelles en ce domaine.

[1] ATILF « Analyse et Traitement Informatique de la Langue Française » UMR 7118 CNRS Nancy Université, 44, avenue de la Libération BP 30687 54063 Nancy cedex ; site Web : http://www.atilf.fr ; courriel : contact@atilf.fr.

- Un *dictionnaire linguistique ou dictionnaire de langue*. Par opposition à une visée encyclopédique, le *Trésor* s'attachera à définir chaque mot par ses caractéristiques linguistiques : sa forme, son sens, ses emplois stylistiques et syntaxiques.
- Un *dictionnaire, œuvre d'une génération*. La création du Centre de Recherche pour un Trésor de la Langue Française coïncida avec les premières utilisations en sciences humaines de moyens mécanographiques et informatiques de documentation. Le TLF est le premier dictionnaire de langue se fondant sur une méthodologie systématique d'analyse des usages effectifs des mots de notre langue à travers l'exploitation d'une vaste base de données textuelles dont la saisie a débuté dès les années 60. Ainsi, un rédacteur ayant à écrire un article se trouvait doté de concordances systématiques de ce mot, triées suivant différents critères : ordre chronologique des sources, ordre alphabétique des contextes gauche et droit ou encore ordre défini selon les constructions syntaxiques propres à chaque partie du discours. Le traitement informatique était assuré par de lourds logiciels, procédant par traitement séquentiel du corpus. Plus tard, au début des années 80, le laboratoire a réalisé une plate-forme de base de données textuelles qui a permis un gain de productivité spectaculaire grâce à la possibilité d'accès direct aux mots du corpus, et, surtout, d'envisager la réalisation d'une première interface utilisateur (1985), avec une exploitation télématique par les moyens de l'époque (terminaux Transpac ou Minitel) : ainsi est née la base textuelle FRANTEXT qui aujourd'hui regroupe plus de 4000 œuvres littéraires (www.atilf.fr/frantext) représentant plus d'un milliard de caractères.

L'évolution des techniques informatiques permet aujourd'hui, à travers l'informatisation d'un dictionnaire, de découvrir des usages nouveaux et des parcours véritablement novateurs qui s'affranchissent des aspects essentiellement séquentiels de la lecture et de la recherche dans les textes imprimés. Ces enjeux sont très vite apparus et, avant même la fin de sa publication papier, l'informatisation du TLF est envisagée. A cette époque, hélas, les techniques informatiques disponibles et les coûts afférents, tant en moyens humains qu'en moyens financiers, ne permettent pas de faire avancer un tel projet. Comme le note Robert Martin[2], dans l'introduction de l'ouvrage *Lexicographie et Informatique* (Piotrowski 1996) « *des chiffrages exorbitants et conséquemment une réponse dilatoire, extrêmement réservée, en mai 1991, de notre partenaire éditorial, les Editions Gallimard, l'impossibilité quasi-absolue de susciter, au départ, l'investissement et la collaboration d'industriels, tout cela faisait apparaître le TLF informatisé comme une chimère* ». Pourtant, en mai 1995 à Nancy, lors d'un colloque international Robert Martin peut affirmer : « *Nous avons la certitude que le TLF sera informatisé ; il est même en bonne voie de l'être* ». Trois éléments sont intervenus dans cette évolution :

- Une collaboration scientifique avec la Bibliothèque Nationale de France qui permit la saisie des huit premiers volumes pour lesquels il n'existait pas d'archives électroniques ;
- Un soutien sans défaut de la direction des Sciences Humaines et Sociales du CNRS et, plus globalement, de la communauté scientifique qui attendait beaucoup d'une telle informatisation, ressource de base pour des recherches futures ;
- Mais surtout, et ce fut là le point déterminant, l'énorme travail réalisé par le service informatique du laboratoire qui, sous la direction de Jacques Dendien, Ingénieur de Recherche CNRS, sut démontrer la faisabilité de la rétroconversion du TLF.

[2] Robert Martin succéda à Bernard Quémada à la direction de l'INaLF.

Lorsqu'en janvier 2001, à la demande du CNRS, j'ai accepté de m'investir dans la direction de ce laboratoire rebaptisé ATILF, j'ai trouvé une équipe motivée, à la tête d'un prototype déjà très élaboré de cette version informatique du TLF. Cet outil demeurait, pourtant, encore inaccessible : il demandait à être finalisé et sa pertinence devait encore être prouvée. Une des premières tâches à laquelle nous nous sommes donc attelés fut d'ouvrir très vite cette première version sur le Web en vue d'un test grandeur nature. Le 5 mars 2002, avec l'accord et le soutien de la direction générale du CNRS, une présentation publique du *TFLi* était organisée. Alors que la moyenne des pages du TLFi consultée en février 2002 était inférieure à 50 par jour, nous sommes passés à plusieurs milliers fin mars 2002 et à plus de 600 000 pages par jour ouvrable aujourd'hui en cumulant les accès direct au TLFi et celles effectuées via le portail lexical du CNRTL (www.cnrtl.fr/portail).

3. Du Trésor de la Langue Française au Trésor de la Langue Française informatisé

Reflet fidèle de la version papier, le TLFi (www.atilf.fr/tlfi) se caractérise, comme le TLF, par la richesse de son matériau et la complexité de sa structure :

- Importance de sa nomenclature : 100 000 mots avec leur étymologie et leur histoire et 270 000 définitions.
- Richesse de chaque article (vedettes, codes grammaticaux, indicateurs sémantiques ou stylistiques, indicateurs de domaines, définitions, exemples référencés…).
- Richesse des 430 000 exemples, tirés de plus de deux siècles de production française.
- Diversité des rubriques : rubrique d'analyse sémantique synchronique (couvrant la période 1789 à nos jours), rubrique « prononciation et orthographe », rubrique étymologie et histoire, rubrique de statistique lexicale et rubrique bibliographie.

Cette version du TLF (Dendien et Pierrel 2003) intègre des accès à très haut niveau de tolérance permettant une insensibilité aux accents et une tolérance aux fautes d'orthographe courantes. De plus, elle offre des accès à partir de formes et non uniquement de lemmes ou de vedettes et propose des procédures d'accès diversifiées pour une consultation humaine.

3.1. Saisie initiale du dictionnaire

La première étape a consisté à réaliser une archive fiable de la totalité des 16 tomes en retranscrivant sur support informatique la totalité du texte au kilomètre. Les huit premiers tomes ayant été composés au plomb, il convenait d'en assurer la saisie. Ce travail a été réalisé grâce à un accord passé avec la Bibliothèque Nationale de France qui a financé l'opération. La saisie a été faite par une société privée à la suite d'un appel d'offres. Des contrôles statistiques, suivant un protocole très rigoureux, ont permis au laboratoire de s'assurer de la qualité de cette saisie.

Les tomes 9 à 16 existaient sous forme de trois formats distincts de photocomposition (trois imprimeurs différents se sont en effet succédé pour réaliser le TLF). Les tomes 9 et 10 ainsi que les tomes 14 à 16 présentaient un état d'archives tout à fait fiable et récupérable. Ils ont été remis dans un format standard grâce à un marché passé avec une société privée spécialisée dans le traitement des archives de photocomposition. L'analyse de l'état des archives des tomes 11 à 13 a malheureusement montré un texte incomplet, désordonné, avec des fragments qui se répétaient curieusement plusieurs fois. Malgré son mauvais état, l'ensemble valait néanmoins la peine d'être récupéré. Remise en ordre et reconstitution des passages manquants ont été réalisées au laboratoire.

3.2. Balisage du contenu du dictionnaire

Les articles du TLF se présentent en deux parties :

- une première partie appelée « synchronie » exposant les différents sens des mots,
- une seconde partie, « diachronie », constituée de plusieurs rubriques consacrées à l'histoire, l'étymologie ou la phonétique.

L'histoire de l'élaboration du TLF s'étant étalée sur une période de plus de trente ans, seule la partie synchronique correspondait à des normes de rédaction relativement stables. Les normes de rédaction de la seconde partie ne se sont stabilisées que fort tardivement, aux environs du tome 11. Avant ce tome, les différentes rubriques sont constituées d'un discours totalement informel qu'il serait vain de vouloir structurer[3].

Les efforts de balisage ont donc porté sur la partie synchronique. On peut y dénombrer environ 40 types d'informations différents (vedettes, codes grammaticaux, indicateurs sémantiques ou stylistiques, indicateurs de domaine ou d'usage, définitions, exemples référencés…). Leur réunion couvre la totalité du texte, permettant ainsi une structuration complète de la synchronie.

Le processus d'informatisation consiste à injecter, dans le texte au kilomètre, des balises textuelles de type XML. Compte tenu de l'importance du TLF (environ 350 millions de caractères) et du nombre d'objets rencontrés, il était hors de question de prétendre réaliser ce travail manuellement. Il fallait donc créer des automates capables de l'effectuer. Mais en raison de la grande variété des types d'objets à reconnaître et de la minceur des indices permettant leur discrimination automatique, il a fallu plus d'un an de travail, marqué par des approches infructueuses, avant d'acquérir la certitude de la faisabilité de l'opération.

A partir du texte initial (cf. figure 1), la reconnaissance des différents objets par les automates a été guidée par les éléments suivants :

LOSANGE, subst. masc.

A. — *GÉOM.* Parallélogramme ayant des côtés égaux et dont les angles ne sont pas droits. Synon. vieilli et littér. *rhombe. Dans un losange deux angles sont aigus et deux sont obtus* (*Ac.* 1878-1935). *Losanges, cercles, ovales et autres figures obtenues avec les compas* (CHAMPFL., *Souffr. profess. Delteil,* 1853, p. 101).

● *En losange.* Suivant la forme d'un losange. *Il lui montra la croix du sud, groupe de quatre étoiles de première et de seconde grandeur, disposées en losange* (VERNE, *Enf. cap. Grant,* t. 1, 1868, p. 237).

Figure 1 : *Texte initial*

- **typographie** : les informations typographiques assez pauvres (gras, italique, petites capitales) sont, à elles seules, bien insuffisantes pour identifier les 40 types d'informations différents mais constituent un indice non négligeable. Une première étape consiste à récupérer et baliser ce type d'information (cf. figure 2).

[3] Aujourd'hui, un projet du laboratoire TFL-Etym (http://www.atilf.fr/tlf-etym/) a pour objectif de restructurer entièrement cette partie diachronique.

<R>LOSANGE,</R> <R>subst. masc.</R><G>A. _</G> <I>GÉOM.</I> <R>Parallélogramme ayant des côtés égaux et dont les angles ne sont pas droits.</R> <R> Synon. vieilli et littér. </R><I>rhombe.</I> <I>Dans un losange deux angles sont aigus et deux sont obtus </I><R>(</R><I>Ac.</I> <R>1878-1935</R><R>).</R> <I>Losanges, cercles, ovales et autres figures obtenues avec les compas </I><R> (</R><C>Champfl.</C><R>,</R><I>Souffr. profess. Delteil,</I><R>1853</R> <R>, p.101</R><R>).</R><G>.</G> <I>Enlosange. </I> <R>Suivant la forme d'un losange. </R> <I>Il lui montra la croix du sud, groupe de quatre étoiles de première et de seconde grandeur, disposées en losange </I><R>(</R><C>Verne</C><R>,</R><I>Enf. cap. Grant, </I><R>t. 1</R> <R>,1868</R><R>, p.237</R><R>).</R><G>B. _</G> <I>

Figure 2 : *Balisage typographique*

- **contenu textuel** : un certain nombre d'objets (par exemple les indications de domaine technique ou de type grammatical, sémantique, stylistique) ont un contenu textuel appartenant à une nomenclature fermée. Leur reconnaissance est donc assez facile, à condition toutefois que cette nomenclature soit connue de manière exhaustive. Dans la pratique, nous l'avons bâti de manière incrémentale : au fur et à mesure que les opérations avançaient, elle était enrichie, avec pour conséquence la nécessité de procéder à un retour-arrière pour corriger ce qui avait déjà été accompli. Les résultats de cette étape ont permis d'enrichir le balisage par un balisage de contenu plus sémantique (cf. figure 3).

<art><ved><mot><R>LOSANGE,</R></mot><cod><R>subst.masc.</R></cod></ved><parah><G>A. _</G> </parah><dom><I>GÉOM.</I> </dom><def n="t"><R>Parallélogramme ayant des côtés égaux et dont les angles ne sont pas droits.</R> </def><syno><R>Synon. vieilli et littér. </R><I>rhombe.</I> </syno><exe n="e"><I>Dans un losange deux angles sont aigus et deux sont obtus</I><R>(</R><pub><I>Ac.</I></pub><dat><R>18781935</R></dat><R>).</R> </exe><exe n="e"><I>Losanges, cercles, ovales et autres figures obtenues avec les compas </I><R>(</R><aut><C>Champfl.</C> </aut><tit><R>, </R><I>Souffr. profess. Delteil,</I> </tit><dat><R>1853 </R></dat><loc><R>, p. 101</R></loc><R>).</R></exe><paraputir> <G>. </G> </paraputir><syntita n="d"><I>En losange.</I> </syntita><def n="t"><R>Suivant la forme d'un losange.</R> </def><exe n="e"><I>Il lui montra la croix du sud, groupe de quatre étoiles de première et de seconde grandeur, disposées en losange </I><R>(</R><aut><C>Verne</C></aut> <tit><R>, </R><I>Enf. cap. Grant, </I><ct><R>t. 1</R></ct></tit><dat><R>, 1868</R></dat><loc><R>, p. 237</R></loc><R>).</R></exe>

Figure 3 : *Enrichissement par balisage de contenu*

- **succession et structuration des éléments :** les différents types d'éléments ne se suivent pas au hasard mais obéissent aux lois des normes de rédaction. Il est donc possible d'identifier certains éléments en fonction du contexte dans lequel ils apparaissent. La dernière étape consiste donc à enrichir ce balisage par un balisage codant cette structure.

Les automates de reconnaissance ont été mis au point progressivement en choisissant des échantillons dans les différents tomes afin de rencontrer les cas de figure les plus variés. La

version N des automates produisait un texte balisé. Les différents types d'erreurs de balisage étaient ensuite classifiés et faisaient l'objet de corrections produisant la version N+1. Après une dizaine d'itérations de ce type, il s'est avéré que les erreurs résiduelles étaient peu nombreuses (taux de réussite des automates de l'ordre de 99.8 %) et toutes atypiques (chaque erreur était due à des circonstances particulières non récurrentes).

Dans la version XML finale, nous avons tenu à conserver la totalité des marqueurs typographiques présents dans le texte initial de manière à conserver une image 100 % fidèle du TLF. En effet, le texte initial comporte un certain nombre de variations typographiques non représentatives d'un type d'objet. Ces variations typographiques sont codées sous forme de balises XML faisant partie intégrante de la structure du dictionnaire et constituent les éléments les plus internes de la structure.

Au total, on peut faire le dénombrement suivant, après validation de l'ensemble des seize tomes :

- nombre de balises typographiques : 17 364 854,
- nombre de balises décrivant la hiérarchie : 1 070 224,
- nombre de balises repérant les objets textuels : 18 178 634, dont 92 997 entrées et 64 346 locutions faisant l'objet de 271 166 définitions et illustrées par 427 493 exemples,
- nombre total de balises XML : 36 613 712,
- niveau de profondeur hiérarchique maximal : 23.

3.3. Développement de ressources complémentaires en vue d'améliorer l'interface

3.3.1. Base lexicale pour une hypernavigation

Les navigateurs Web les plus répandus permettent, lorsque l'on procède à un double clic sur un mot quelconque d'une page Internet, de récupérer ce mot grâce à une procédure informatique spécifique. Il est alors possible de passer le mot ainsi sélectionné à une autre application. Tous les mots d'une page Internet peuvent donc être considérés comme autant d'hyperliens virtuels. Cette possibilité est mise à profit dans certains dictionnaires en ligne comme par exemple les dictionnaires de Cambridge (http://dictionary.cambridge.org/). L'effet d'un double clic sur un mot quelconque d'un article est de provoquer sa recherche dans les autres articles du même dictionnaire. Cette intéressante possibilité a été généralisée dans le TLF à toutes les applications développées par l'ATILF.

Mais un problème se pose si un utilisateur navigant dans un dictionnaire sélectionne une forme flexionnelle. Il est vraisemblable alors que sa véritable intention est de déclencher une hypernavigation sur le lemme et non pas sur la forme flexionnelle. Dans le cas où la forme flexionnelle est ambiguë (non-unicité du lemme), on se trouve confronté au problème classique de la désambiguïsation bien connu en traitement automatique des langues.

Dans le cas des applications de l'ATILF, toute application peut à la fois être la source (celle d'où provient le mot sélectionné par double clic) et la cible (celle qui traite le mot associé au double clic) d'une hypernavigation. Lors du double clic sur un mot, nous ne transmettons que la forme à l'application-cible. C'est donc l'application-cible qui procède éventuellement à la lemmatisation de la forme (malheureusement, en absence de contexte, plusieurs lemmes sont parfois envisageables).

En cas d'un double clic dans une application quelconque, nous proposons un menu surgissant (cf. figure 4) qui permet de déclencher une recherche dans différentes bases textuelles ou lexicales :

- le TLF,
- la quatrième édition du dictionnaire de l'Académie française (1762),
- la huitième édition du dictionnaire de l'Académie française (1935),
- la neuvième édition du dictionnaire de l'Académie française (édition actuelle, en cours. A ce jour, les académiciens en sont à la lettre P),
- une base dite de « connaissance lexicale » qui permet d'expliquer à l'utilisateur l'origine de la forme qu'il a sélectionnée. Par exemple, s'il sélectionne la forme `éditions`, il lui sera expliqué qu'il s'agit soit d'un substantif féminin pluriel ayant pour lemme `édition`, soit de la première personne du pluriel de l'imparfait de l'indicatif ou du présent du subjonctif ayant pour lemme `éditer`,
- la base de données FRANTEXT. On s'imaginera sans peine la richesse apportée par la dualité FRANTEXT TLF : dans le sens FRANTEXT vers TLF, le TLF est une aide à la compréhension des textes, et dans le sens TLF vers FRANTEXT, FRANTEXT vient compléter puissamment le jeu des exemples cités dans le TLF ,
- la Base Historique du Vocabulaire Français (dictionnaire des datations de l'origine des mots réalisé par notre laboratoire).

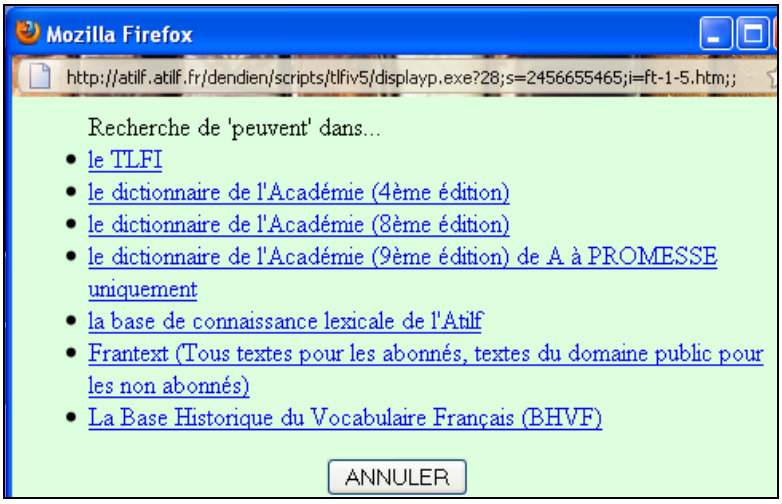


Figure 4 : *Fenêtre d'hypernavigation*

3.3.2. Nature de la base lexicale créée pour l'hypernavigation

L'application-cible d'une hypernavigation récupère la forme sélectionnée et a éventuellement besoin de connaître l'ensemble des lemmes pouvant correspondre à cette forme. Afin de permettre la lemmatisation rapide des formes, nous avons développé une base de données lexicales permettant de lier les formes à leur(s) lemme(s) ainsi qu'aux informations grammaticales associées (mode, temps, personne pour les verbes, genre et nombre pour les substantifs ou adjectifs).

Une telle ressource n'est pas en soi une réalisation bien originale, mais la version que nous avons réalisée est issue du TLF : sa nomenclature est égale à celle du TLF. Les avantages sont que, d'une part, le succès d'une hypernavigation vers le TLF est ipso facto garanti, et que,

d'autre part, nous disposons là d'une ressource très riche, réutilisable par toute application opérant sur des textes littéraires et nécessitant des opérations de flexion et de lemmatisation.

L'ensemble des informations de cette base (lemmes, formes, informations associées) est codé dans un format XML dûment validé, ce qui facilite sa réutilisation dans le cadre d'autres applications. Il est aujourd'hui distribué sous forme d'un lexique ouvert des formes fléchies du français MORPHALOU, accessible à l'adresse www.cnrtl.fr/lexiques/morphalou.

3.3.3. Accessibilité de la base lexicale via un serveur

Il est possible d'exploiter directement la base lexicale que nous venons de décrire sans qu'il soit pour autant nécessaire de réaliser un développement informatique spécifique. A l'instar de ce qui a été fait pour l'interrogation du TLF, nous avons l'avons dotée de la possibilité d'un accès à distance en soumettant au serveur des requêtes de type XML ou requête Web. Ainsi la requête www.cnrtl.fr/morphologie/sussiez fournit en résultats l'analyse morphologique des diverses forme du verbe `savoir`.

3.3.4. Base de données phonétiques

Problèmes soulevés par les utilisateurs des dictionnaires électroniques

L'utilisation d'un dictionnaire électronique, en apportant des possibilités de consultation transversale, apporte indéniablement un progrès extraordinaire par rapport à la version papier du même dictionnaire. Le développeur d'interface d'interrogation de dictionnaires électroniques est donc tout naturellement tenté de porter ses efforts sur la consultation transversale au détriment de l'utilisation la plus simple : la recherche d'un mot. Nous allons montrer que c'est là une grave erreur.

Depuis plusieurs années déjà, l'ATILF propose en libre accès aux internautes, avec le TLF et les deux dernières éditions du dictionnaire de l'Académie, plusieurs grands dictionnaires informatisés. Nos serveurs sont actuellement sollicités par plusieurs centaines de milliers de requêtes quotidiennes.

L'analyse des demandes permet de faire ressortir les éléments suivants :

- pour l'utilisateur « grand public », la fonction essentielle d'un dictionnaire est de vérifier le sens ou l'orthographe d'un mot donné. La proportion d'utilisateurs procédant à des recherches transversales est infime,
- si l'on propose une simple case blanche à remplir pour rechercher un mot (comme c'est le cas dans la quasi-totalité des dictionnaires électroniques proposés sur le Web) on constate que l'utilisateur reste sans réponse, après plusieurs tentatives, dans une proportion assez considérable (de l'ordre de 10 à 15 %), bien que le mot recherché figure bel et bien dans le dictionnaire.

Ce constat est cuisant. Le grand public fait donc fi de l'amélioration apportée par la recherche transversale et se heurte, d'autre part, à des difficultés lorsqu'il utilise le dictionnaire dans la finalité sa plus simple : permettre la recherche d'un mot.

Après avoir fait porter nos efforts sur les recherches transversales, nous avons donc décidé de trouver une solution efficace au problème de la recherche d'un mot.

Les échecs constatés venaient en grande partie d'un fait très simple : il est ridicule de demander à l'utilisateur de taper l'orthographe exacte du mot recherché, quand c'est précisément cette orthographe qu'il recherche. Un dictionnaire papier apporte une solution acceptable à ce problème : son lecteur émet un certain nombre d'hypothèses sur l'orthographe recherchée et essaie de les valider en feuilletant le dictionnaire. Cette approche peut être

reproduite dans le cas d'un dictionnaire électronique, en proposant des listes défilantes de mots dans lesquelles l'utilisateur va chercher son bonheur. Même si elle déplait à l'utilisateur pressé, par son caractère fastidieux, cette méthode, que nous proposons aussi pour le TLFi, a l'avantage de proposer un parcours “ ludique ” : qui, cherchant un mot dans un dictionnaire, n'a pas eu le regard attiré par un autre mot dont il consulte l'article ?

Une approche plus intéressante pour éviter les fautes consiste à doter le dictionnaire électronique d'une certaine tolérance aux fautes, à l’exemple de ce qui est fait dans le domaine de la correction orthographique des éditeurs de texte.

Ceci reste cependant bien insuffisant. En effet, l'expérience du TLFi nous a montré que l'utilisateur peut formuler des requêtes bien inattendues :

- absence systématique d'accents ou de cédilles (pratique liée au contexte informatique ?),
- formes flexionnelles de verbe, de noms ou d'adjectifs assorties d'éventuelles fautes d'orthographe. Ce comportement se retrouve à la fois chez des utilisateurs à niveau culturel limité (certains ignorent qu'un dictionnaire donne une classification par lemmes), chez des utilisateurs cultivés (mais le simple fait d'utiliser un objet électronique leur fait oublier leurs réflexes d'utilisation d'un dictionnaire), ou encore chez des utilisateurs peu familiers de la langue française (ils ignorent tout simplement que ce qu'ils recherchent est une forme flexionnelle),
- expressions (par exemple `monnaie de singe`). On peut conjecturer que l'utilisateur, même lorsqu'il sait qu'avec un dictionnaire papier il lui faudrait sans doute chercher dans les entrées `singe` ou `monnaie`, estime que la moindre des choses est que le dictionnaire électronique auquel il s'adresse lui épargne ce genre de tracas,
- membres entiers de phrase (par exemple `craindre qu'il soit`). Dans ce cas, l'usage attendu n'est plus de contrôler l'orthographe d'un mot mais de contrôler l'usage de tournures syntaxiques.

La fréquence de ces comportements est récurrente. Certains d'entre eux correspondent à l'idée qu'un dictionnaire électronique doit offrir des services supérieurs à ceux d'un dictionnaire papier. Une telle attente nous semble légitime. D'autres résultent d'une méconnaissance de la langue. Dans ce dernier cas, tout doit être tenté pour satisfaire l'utilisateur qui a fait l'effort de consultation.

Nécessité d'un traitement phonétique

La prise en compte des problèmes précédemment cités nécessite la mise en œuvre de différentes mesures :

a) une correction orthographique prenant en compte les problèmes les plus récurrents (absence d'accentuation ou accentuation erronée, problème du redoublement de consonnes, confusion entre les lettres I et Y, présence de H muets à l'intérieur des mots, etc.),

b) la prise en compte du fait que l'utilisateur a peut-être fourni une forme flexionnelle au lieu d'un lemme : il sera nécessaire de procéder aux lemmatisations nécessaires,

c) la nécessité de donner à l'utilisateur une issue de secours en lui permettant de fournir un équivalent phonétique de ce qu'il cherche,

d) l’acceptation que la recherche de l'utilisateur puisse porter sur les entrées du dictionnaire bien entendu, mais aussi sur tout le texte du dictionnaire (par exemple il est fort possible que l'utilisateur qui désire contrôler une tournure syntaxique en trouve confirmation au détour d'un exemple cité dans le dictionnaire).

Le point **a)** peut se résoudre avec les techniques ordinaires de correction orthographique. Nous ne nous y attarderons pas.

Le point **b)** peut se résoudre avec l'assistance de la base de données lexicales que nous avons décrite plus haut (cf. paragraphe 4.3.2) à propos de l'hypernavigation. Il est donc d'emblée résolu.

Le point **d**) est lui aussi facilement résolu dans le TLFi car la structure même du dictionnaire (base semi-structurée codée en XML) permet tout à la fois un accès via la structure du dictionnaire et une recherche plein-texte.

Le point **c)** nécessite, en revanche, la création de ressources permettant des traitements phonétiques. Ces ressources doivent permettre de transformer la donnée de l'utilisateur en données phonétiques, et à l'inverse de passer des données phonétiques aux mots pour retrouver ce que l'utilisateur avait peut-être voulu dire.

Pour créer une telle ressource, nous sommes partis de la nomenclature de la base créée pour l'hypernavigation et avons cherché à associer à chaque graphie de la base sa prononciation.

Une telle base de données permet la prise en compte des phénomènes d'homophonie, mais part du pré-supposé que le mot dont on recherche l'homophone (ou les homophones) est orthographié correctement. Si l'on prend en compte que l'utilisateur ne respecte pas l'orthographe, cette hypothèse ne sera que rarement respectée. Il est donc nécessaire de compenser cette déficience par un mécanisme complémentaire permettant le passage vers la forme phonétique. Le mécanisme que nous avons choisi est un logiciel du genre TTS (Text To Speech) capable de prendre le relais lorsque la base de données phonétiques ne fournit pas de réponse.

Processus de création de la base de données phonétiques

Le TLF fournit en principe la présentation phonétique de ses entrées. Malheureusement, ceci n'est systématique qu'à partir de la tranche alphabétique E. D'autre part, il comporte un grand nombre d'entrées secondaires (mots ne faisant pas partie d'une entrée principale mais cités comme dérivés d'entrées principales). Dans ce cas, la représentation phonétique d'un mot ou d'une locution est rarement donnée. Le TLF est donc une source de données insuffisante pour constituer une base de données phonétiques. Nous avons réussi à combler la presque totalité des lacunes avec d'autres sources disponibles sur Internet.

En soumettant au logiciel TTS (cf. paragraphe précédent) l'ensemble de la nomenclature du TLF, nous avons pu disposer (du moins en ce qui concerne les lemmes) de plusieurs versions phonétiques d'un même mot (en provenance du logiciel TTS, du TLF ou de sources extérieures). La comparaison de ces différentes versions a fait émerger plusieurs milliers de contradictions qui ont été tranchées en mode manuel.

Il restait à effectuer la validation de la phonétique des formes flexionnelles pour lesquelles la seule source disponible était le résultat du logiciel TTS. Fort heureusement, à chaque règle morphologique permettant de passer d'un lemme à une forme flexionnelle, il est possible d'associer une règle permettant d'élaborer la forme phonétique de la forme flexionnelle par assemblage de la forme phonétique du lemme et de la forme phonétique de la désinence. La quasi-totalité de la phonétique des formes flexionnelles (à l'exclusion des formes flexionnelles des verbes du troisième groupe) a pu être ainsi contrôlée et éventuellement corrigée.

3.3.5. Traitements des données fournies par l'utilisateur : études de cas.

Nous avons exposé plus haut que ces traitements sont de trois ordres (correction orthographique, traitement phonétique, lemmatisation).

Prenons le cas d'un utilisateur ayant tapé « `jenero` ». On trouve dans cet exemple la conjonction de trois types de difficultés (absence d'accent, mot tapé phonétiquement, fourniture d'une forme flexionnelle).

La première phase du traitement consiste à procéder à une correction orthographique portant sur l'accentuation. Le logiciel va donc envisager les hypothèses `jenero, jenéro, jénero, jénéro`. À toutes les hypothèses il va appliquer le mécanisme de passage vers une représentation phonétique et donc chercher (sans succès au demeurant) chacune des quatre formes envisagées dans une base de données phonétiques. Cependant, le mécanisme TTS va générer quatre formes phonétiques qui, chacune, seront ensuite recherchées dans la base de données phonétiques (cette fois pour un passage forme phonétique vers graphie). Les trois premières se traduiront par une non-réponse, mais, fort heureusement, la base de données fournira la solution `généraux` pour la quatrième. Les traitements seront enfin complétés par une phase de lemmatisation donnant le lemme `général`. L'utilisateur obtiendra cette réponse. Dans le cas où il n'y a pas unicité de la réponse, les différents résultats seront proposés à l'utilisateur : à lui de faire son choix, cette solution étant de loin préférable à une absence totale de réponse.

Dans le cas où l'utilisateur introduit non pas un mot unique, mais une suite de plusieurs mots, la séquence des traitements exposés ci-dessus sera appliquée à chacun des mots de la suite.

Supposons maintenant que l'on trouve n'importe où dans le TLF une séquence de mots m_1 m_2 m_3, telle que m_1 soit une des hypothèses retenues pour `saigné`, m_2 une des hypothèses retenues pour *a* et m_3 une des hypothèses retenues pour `blanc`. Une telle séquence, alors, est considérée comme une réponse pertinente à la demande de l'utilisateur. En l'occurrence, seule la séquence `saigner à blanc` sera effectivement trouvée dans le TLF.

Le traitement mot par mot décrit ci-dessus est complété par un traitement phonétique global de la séquence. Ce mécanisme est particulièrement utile pour le traitement des mots composés dont on ne sait trop s'ils doivent être écrits attachés ou non (*anticapitalisme*, *paranormal*, etc.). Il est amusant de constater qu'il dote également le logiciel du TLF de la possibilité d'appréhender les calembours (rechercher « `aile et faon` » ou « `sauces y sont` » `!`).

3.4. Exemples de recherches dans le TLFi

On peut trouver à l'adresse www.tlfi.fr une présentation et des démonstrations des modes de recherche offerts dans le TLFi, mais la meilleure façon de se rendre compte de l'intérêt d'une telle transformation du TLF en document numérique consiste soit à accéder au Cédérom du TLFi (ATILF, 2004), soit à se connecter directement à l'adresse : www.atilf.fr/tlfi. Trois principaux types d'accès sont alors proposés : la recherche d'un mot, la recherche assistée et la recherche complexe.

3.4.1. Recherche d'un mot

Cette recherche permet un accès à un mot à travers un système de correction et de lemmatisation automatique (forcée ou non) : ainsi, en introduisant la recherche de la forme *etique* (sans accent), on accède aux deux articles correspondant aux mots *étique* ou *éthique ;* de même un accès à partir de la forme *sussiez* permet d'obtenir automatiquement l'article *savoir*. Elle donne aussi la possibilité d'obtention directe des définitions et conditions d'usage d'une unité lexicale ne disposant pas d'un traitement lexicographique autonome (on accède par exemple au substantif masculin *trompette*, traité dans un super-article *trompette* qui

englobe tant le masculin que le féminin, à travers la requête « le trompette ») ou d'une expression comme *battre la mesure*, en focalisant la réponse sur l'élément pertinent demandé et en offrant la possibilité, à l'aide d'une sorte de « stabilo boss » électronique, de surligner tel ou tel objet textuel. Ici, par exemple, la définition :

```
Objets de la recherche : 1Paragraphe1
```
H TROMPETTE, subst.

II. —*Subst. masc.* Personne qui joue de la trompette.
1A. —Soldat chargé d'exécuter les sonneries. *Le trompette de l'escadron, d'un régiment de cavalerie. Tu seras capitaine, avec une nuée de trompettes courant et sonnant devant toi* (HUGO, *Légende*, t. 3, 1877, p. 390).1
—*Loc. fam., vieilli. Il est bon cheval de trompette*. Il ne se laisse ni effrayer, ni intimider. *Son air, un air de bon cheval de trompette qui ne craignait pas le bruit* (A. DAUDET, *Tartarin de T.*, 1872, p. 13).
B. —Musicien jouant dans une fanfare, un orchestre. Synon. *trompettiste* (*infra* dér.). *Le trompette noir du dancing* (BEAUVOIR, *Mandarins*, 1954, p. 306).

3.4.2. Recherche assistée

Ce second type d'accès permet d'établir la liste des composés comportant un élément donné : ainsi, en demandant les composés contenant le lexème *queue*, on obtient 35 réponses dont :

COURTE-QUEUE, adj. et subst.
DEMI-QUEUE, subst. fém.
HOCHEQUEUE, HOCHE-QUEUE, subst. masc.
PAILLE-EN-CUL, PAILLE-EN-QUEUE, subst. masc.
PORTE-QUEUE, subst. masc.
Etc.

La recherche assistée permet aussi de rechercher *« les verbes qui, en marine, concernent le maniement des voiles »*. Il suffit de préciser que l'on recherche dans la *classe des verbes* ceux qui, appartenant au *domaine* notionnel de la *marine*, correspondent à une *définition* incluant une forme flexionnelle (singulier ou pluriel) du mot `voile`, soit dans une structure plus compacte : [code grammatical : *verbe* ; domaine : *marine* ; type d'objet : *définition*, contenu : &m*voile*[4]]. Voici un extrait des 61 réponses que l'on obtient :

[4] &ms*ubs* permet de tester toutes les formes d'un *substantif*, de même que &c*verbe* toutes les formes d'un *verbe*.

ABRIER, ABREYER, verbe trans.

3Empêcher le vent, en l'interceptant, de passer jusqu'à (une autre voile) :3

AGRÉER[2], verbe trans.

3,,Préparer ou travailler à la garniture, aux agrès d'un bâtiment, fourrer les dormans, estroper les poulies, garnir voiles, vergues, etc. : `` (WILL. 1831) :3

AMURER, verbe.

3Fixer l'amure d'une voile pour l'orienter selon le vent :3

ETC……

Autre exemple : pour l'ensemble des mots dont la définition contient le nom `liberté` [type d'objet : définition, contenu : &mliberté], on obtient 306 réponses dont :

1Définition1

ABUSER, verbe trans.

1Exagérer dans l'usage d'une possibilité, d'une liberté :1

AFFRANCHI, IE, part. passé, adj. et subst.

1(Celui) à qui on a donné la liberté.1

AISE[1], subst. fém.

1Grande liberté.1

ALIÉNANT, ANTE, part. prés. et adj.

1Qui prive l'homme de son humanité, de sa liberté :1

Etc…..

3.4.3. Recherche complexe

Les interrogations possibles au sein de ce dictionnaire peuvent prendre des formes encore plus complexes. Ainsi, il est possible de répondre à la requête suivante : « *Quels sont les substantifs empruntés à une langue étrangère (non précisée) et qui sont employés dans le domaine de l'art culinaire ?* ». Il convient pour cela d'utiliser l'onglet « recherche complexe » et de préciser :

Objet 1 : type "*Entrée*",

Objet 2 : type "Code grammatical", contenu "substantif", lien "inclus dans l'objet 1",

Objet 3 : type "Domaine technique", contenu "art culinaire", lien "dépendant de l'objet 1",

Objet 4 : type "Langue empruntée", lien "dépendant de l'objet 1".

Le lien "inclus dans l'objet 1" de l'objet 2 exprime que l'entrée est un substantif, le lien "dépendant de l'objet 1" de l'objet 3 exprime que l'indication de domaine technique est dans la portée de l'objet 1, et le lien "dépendant de l'objet 1" de l'objet 4 exprime que l'objet est dans l'article dont l'entrée est l'objet 1.

Une telle interrogation nous fournit 42 résultats, parmi lesquels :

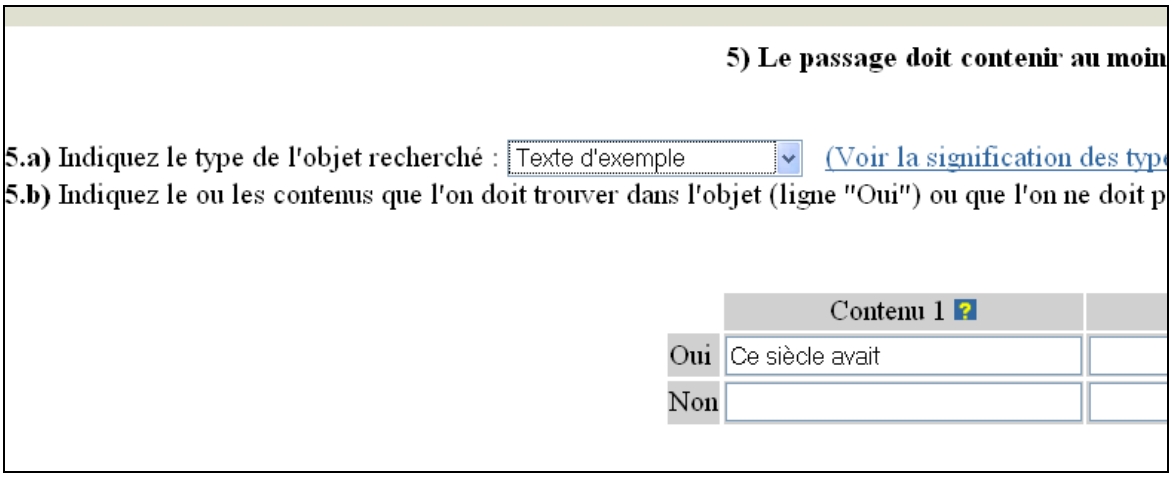
Objets de la recherche : 1Entrée1 2Code grammatical2 3Domaine technique3 4Langue empruntée4

BOR(T)SCH, subst. masc.
1 4Empr. au russe4
CARAMEL, subst. masc.
2 4Empr. à l'esp.4
CAVIAR, subst. masc.
3 4Empr. au vénitien4
CONDIMENT, subst. masc.
4 4Empr. au lat. class.4
ESSENCE[3], subst. fém.
5 4Empr. au lat. class.4
ESTOUFFADE, subst. fém.
6 4Empr. à l'ital.4
GANACHE, subst. fém.
7 4Empr. à l'ital.4

3.4.4. Recherche de citations

Un dernier usage souvent exploité dans le TLF est la recherche de citation. En effet le TLF contient environ 430 000 exemples d'usage avec leur référence. Ces exemples ayant été soigneusemant choisis, ils correspondent à un ensemble de citations potentielles.

Ainsi, si un utilisateur souhaite retrouver une citation dont il n'a qu'un souvenir partiel, il peut indiquer dans la fenêtre d'accueil qu'il recherche dans un *texte d'exemple* ce dont il se rappelle, par exemple `ce siècle avait` :



Il obtient alors les articles correspondant à sa demande soit sous forme simplifiée (cf. Figure 5) soit sous forme étendue (cf. Figure 6) et retrouve ainsi la citation complète avec ses références précises.

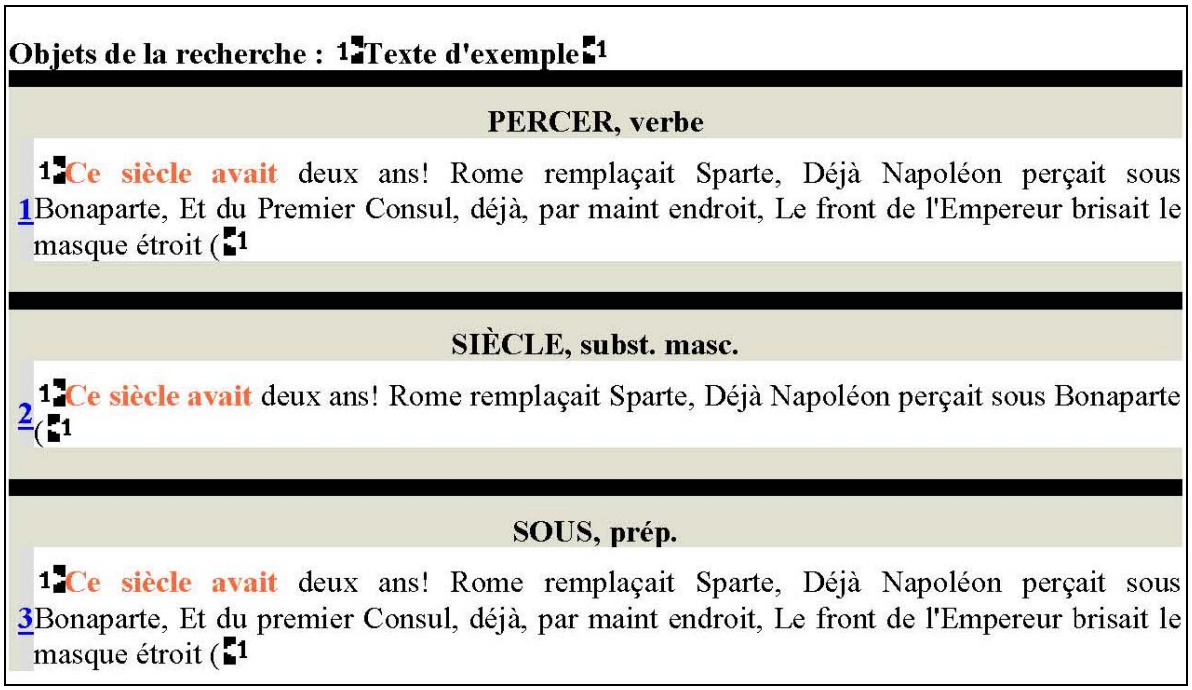


Figure 5 : *Résultats simplifiés*

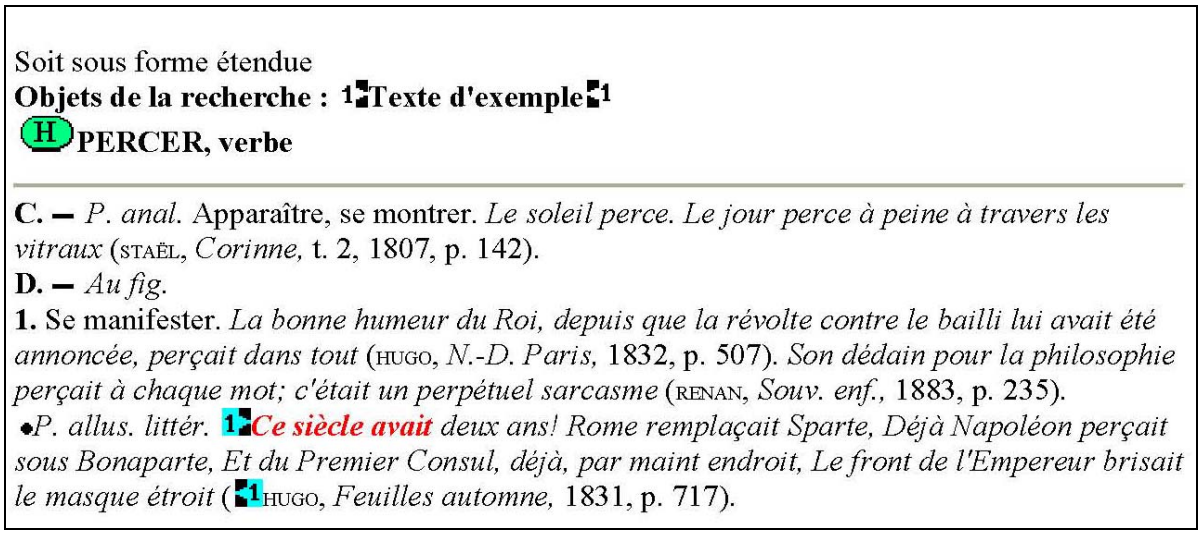


Figure 6 : *Résultat étendu*

4. Conclusion : L'informatique, un vecteur de valorisation des recherches en lexicographie

Concernant la langue française, le TLFi, grâce à la richesse de son contenu entièrement encodé en XML, a ouvert des perspectives intéressantes. Le TLF a eu pendant longtemps la réputation tenace d'être un dictionnaire réservé à une élite. Cette perception du TLF pouvait s'expliquer par au moins trois caractéristiques de sa version papier :

- Sa taille : 16 volumes de plus de 1 000 pages chacun,
- Sa richesse de description qui parfois nuisait à sa lecture, au moins pour les articles les plus lourds : l'article `aimer` se développe ainsi sur 12 pages (soit 24 colonnes) et il n'est pas toujours aisé pour un non-spécialiste d'appréhender cette information très riche,

- Son coût, environ 1 500 euros, qui ne le rendait pas facilement accessible à tous.

S'il a su se positionner comme une référence en lexicographie française, la diffusion de sa version papier s'est néanmoins limitée à quelques milliers d'exemplaires au sein d'une intelligentsia somme toute limitée.

Sa version informatique sous forme de Cédérom (environ 15 000 exemplaires vendus en moins de 4 ans) ou de ressources librement accessibles sur le Web a rencontré un succès important tant auprès du grand public que des utilisateurs universitaires ou des professionnels de la langue : plus de 600 000 requêtes par jour.

Sa version Web (www.atilf.fr/tlfi) fait l'objet d'environ 300 000 connexions quotidiennes en provenance de tous les continents, et il est référencé par d'innombrables sources. La notoriété qu'il a acquise en fait un outil de promotion appréciable de la langue française.

Son intégration plus récente encore dans le portail lexical du CNRTL et ses interconnexions avec d'autres types de ressources sur le vocabulaire français le positionnent au cœur d'un ensemble de ressources sur la langue française où il joue un rôle actif et prépondérant, démontrant ainsi que sa réputation élitiste est devenue largement injustifiée. Sa diffusion au sein du portail lexical du CNRTL (www.cnrtl.fr/portail) en fait aujourd'hui l'un des dictionnaires les plus exploités sur le Web : d'environ 350 000 requêtes par jour (cf. Figure 7).

Figure 7 : *Statistique d'usage du portail lexical du CNRTL*

Notons pour conclure que le partage d'une telle version informatisée d'une production scientifique de référence offre aujourd'hui des modes nouveaux de valorisation de ressources ou de résultats de recherche. Au-delà du seul monde universitaire, ces techniques permettent de mettre à disposition de l'ensemble de la société nos résultats de recherche. On peut, pour s'en convaincre, analyser les commentaires apparaissant sur le Web dans de multiples sites institutionnels ou professionnels.

La généralisation de telles exploitations et valorisations de versions électroniques est ainsi en train de modifier notablement les modes de travail et d'échanges scientifiques au sein des communautés de recherche SHS.

5. Bibliographie

ATILF *Trésor de la Langue Française informatisé*, CNRS Editions, 591 p. et CD du texte intégral, Version PC, ISBN 2-271-06273-X, 2004, Version Mac OX X, ISBN 2-271-06365-5, 2005.

Dendien J., Pierrel J.M. Le Trésor de la Langue Française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence, *TAL* Vol 44 – n° 2/2003, Hermès Sciences Edition, p. 11-37.

Imbs P., Quemada B. (dir.). Trésor de la Langue Française. Dictionnaire de la langue du XIXe et du XXe siècle (1789—1960), 16 vol., Paris, Éditions du CNRS/Gallimard, 1971–1994.

Namer F. *Morphologie, lexique et traitement automatique des langues : l'analyseur DériF*, Collection TIC et sciences cognitives, Hermès Sciences Edition, 2009 - 448p.

Pierrel J.M., Buchi E. Research and Resource Enhancement in French Lexicography: the ATILF Laboratory's computerized resources in *Lexicography in Italy and in Europe*, Silvia Bruti, Roberta Cella and Marina Foschi Albert, editors, Cambridge Scholars Publishing, p. 79-118, 2009

Piotrowski D. (sous la direction de), *Lexicographie et Informatique : autour de l'informatisation du Trésor de la Langue Française* , INaLF, Didier Erudition, Paris, 1996

Quemada B. (dir.), *Matériaux pour l'histoire du vocabulaire français. Datations et documents lexicographiques*, 2e série, 48 vol., Besançon/Paris, Centre d'Étude du Vocabulaire Français/Didier/Klincksieck, 1970-1998.