Structuration et usage de ressources lexicales institutionnelles sur le français

Jean-Marie Pierrel

Chapitre 4

Structuration et usage de ressources lexicales institutionnelles sur le français

Jean-Marie Pierrel
Université de Lorraine, ATILF, UMR 7118, Nancy, F-54000, France
CNRS, ATILF, UMR 7118, Nancy, F-54000, France

Introduction

Comme ce fut souvent souligné, dès que l'on s'intéresse à la langue, que cela soit pour un usage strictement humain ou pour une intégration dans une chaîne de traitement automatique, les informations lexicales, liées aux mots de la langue, occupent une importance primordiale (Laporte 97 ; Pierrel 2000). Pourtant, pour le français comme pour bien d'autres langues, il convient de noter qu'à ce jour il n'existe pas de ressources lexicales optimales adaptées tout autant à un usage humain qu'à une intégration dans une chaîne de traitement automatique.

Néanmoins, depuis la seconde moitié du 20e siècle, de nombreuses contributions institutionnelles à la lexicographie française se sont développées qui, pour la plupart, ont donné lieu à une version informatisée. Parmi elles, on peut noter entre autres : le *Trésor de la langue française* (TLF) développé par le CNRS à Nancy (Imbs & Quémada 1971-1994) et sa version informatisée[1] (Dendien & Pierrel 2003), la 9e édition du *Dictionnaire de l'académie française* (Académie 2005) et sa version informatisée[2], le *Dictionnaire du moyen français*[3] (Martin, Gerner & Souvay 2007), la *Base de données lexicale panfracophone*[4] (BDLP) élaborée à l'Université Laval de Québec (Poirier 2005) avec le soutien de l'AUF (Agence universitaire de la francophonie).

Ces études en lexicographie française se sont enrichies, au-delà de la rédaction des dictionnaires, par un effort particulier de valorisation informatique. Plus récemment, le portail lexical du CNRTL[5] (Centre national de ressources lexicales et textuelles) mis en place au sein de l'ATILF regroupe en son sein ces principales ressources lexicales institutionnelles sur le français en y adjoignant certains outils ou données complémentaires tels la version informatisée du *Ducange* de l'École des Chartes, l'accès à un concordancier s'appuyant sur les textes libres de droits de Frantext[6] (Bernard, Lecomte, Dendien & Pierrel 2002) ou une représentation en trois dimensions de la proximité entre les mots, la *proxémie*, mise au point au CLLE-ERSS (Gaume 2004).

Ce choix d'interconnecter diverses ressources lexicales institutionnelles a provoqué sur le plan des études lexicales une véritable révolution (Pierrel 2008) qui fit de l'informatique un outil indispensable pour :

a) étudier le lexique et ses propriétés à travers l'exploitation intelligente de diverses ressources informatisées ;

b) structurer et normaliser les connaissances lexicales et lexicographiques ;

c) valoriser, partager et mutualiser les résultats de la recherche sur le lexique de notre langue, trop souvent encore dispersés.

---

[1] www.atilf.fr/tlfi
[2] http://www.academie-francaise.fr/dictionnaire/
[3] http://www.atilf.fr/dmf/
[4] http://www.bdlp.org/
[5] www.cnrtl.fr/portail
[6] Frantext : base de données textuelles de l'ATILF riche de plus de 4 000 œuvres, accessible à l'adresse www.atilf.fr/frantext.

Ces ressources, librement accessibles sur le Web, ont rencontré un succès important tant auprès du grand public que des utilisateurs universitaires ou des professionnels de la langue. Référencées par d'innombrables sources, elles font l'objet de plusieurs centaines de milliers de connexions quotidiennes[7] en provenance de tous les continents, devenant ainsi une sorte de métadictionnaire incontournable et un outil de promotion appréciable de la langue française. Particulièrement bien adaptées à un usage par un humain, ces ressources demeurent néanmoins rétives à une intégration dans des chaînes de traitement automatique et c'est cette limitation qui conduisit à définir un nouveau projet RELIEF (*Ressources lexicales informatisées d'envergure pour le français*) (Lux-Podogolla & Polguère 2011) dont l'objectif vise, précisément, le développement d'une modélisation informatisée à large couverture du lexique français, le *Réseau lexical du français* ou RLF, exploitable dans un contexte de traitement automatique de la langue

1. Dictionnaires papier versus ressources informatiques

Dans ce chapitre, après avoir présenté les caractéristiques des dictionnaires papier et des dictionnaires informatisés, ainsi que des processus ayant permis de passer de l'un à l'autre, nous analyserons les usages nouveaux offerts par les dictionnaires informatisés avant de présenter les efforts actuels pour modéliser des ressources lexicales mieux adaptées au traitement automatique des langues.

1.1. Dictionnaires institutionnels de référence

Dans le domaine des dictionnaires institutionnels de référence rédigés au cours des dernières décennies, deux grands dictionnaires français, représentant chacun un type de dictionnaire de langue spécifique, méritent d'être mis en exergue : le *Dictionnaire de l'Académie française*, dictionnaire normatif dont l'objectif est de décrire le bon usage de la langue, d'une part, et le *Trésor de la langue française*, dictionnaire représentatif de l'usage effectif de la langue, d'autre part.

La 9e édition du *Dictionnaire de l'Académie française* se situe dans l'héritage d'une longue tradition depuis sa première édition de 1694. Elle est le résultat de la mission assignée à l'Académie dès l'origine[8], à savoir *fixer la langue française, lui donner des règles, la rendre pure et compréhensible par tous*. La première édition de celui-ci fut publiée en 1694, les suivantes en 1718, 1740, 1762, 1798, 1835, 1878, 1932-1935. La neuvième édition, dont la publication a débuté en 1992, est en cours d'achèvement au sein du service du dictionnaire, qui soumet ses travaux à la commission du dictionnaire composée de douze académiciens et chargée de la révision de la précédente édition et de l'élaboration définitive de la nouvelle édition. Le premier tome de la neuvième édition du *Dictionnaire de l'Académie française* (de *A* à *Enzyme*) comporte 14 024 mots, dont 5 500 mots nouveaux par rapport à l'édition précédente. Le deuxième tome (de *Éocène* à *Mappemonde*) comporte environ 11 500 mots, dont 4 000 mots nouveaux. Le troisième tome (de *Maquereau* à *Quotité*) comporte 9 860 mots, dont 3 828 mots nouveaux. La matière du quatrième tome est publiée en fascicules dans les « Documents administratifs » du *Journal officiel,* au fur et à mesure de l'avancement des travaux. Quoi de mieux pour préciser les objectifs de ce dictionnaire que de citer la préface de ce dictionnaire[9], rédigée par son ancien secrétaire perpétuel Maurice Druon : « *Nous ne donnons entrée, parmi les termes techniques, qu'à ceux qui, du langage du spécialiste, sont passés par nécessité dans le langage courant, et peuvent donc être tenus pour réellement usuels. Nous ne faisons place aux mots étrangers qu'autant qu'ils sont vraiment installés dans l'usage, et qu'il n'existe pas déjà un honnête mot français pour désigner la même chose ou exprimer la même idée. Nous sommes d'ailleurs plus accueillants qu'on ne le prétend, considérant que la langue est moins mena-*

[7] http://www.cnrtl.fr/aide/stat/

[8] « La principale fonction de l'Académie sera de travailler avec tout le soin et toute la diligence possibles à donner des règles certaines à notre langue et à la rendre pure, éloquente et capable de traiter les arts et les sciences. » Statuts et règlement de l'Académie française, 1635.

[9] http://www.academie-francaise.fr/dictionnaire/index.html

cée par l'extension du vocabulaire que par la détérioration de la syntaxe. Nous sommes assez rigoureux à l'égard des néologismes, dont beaucoup ne doivent leur apparition qu'à l'ignorance ou l'oubli de bons termes existant depuis fort longtemps ; nous sommes généralement impitoyables s'ils sont formés d'une manière qui insulte au génie de la langue. Les extensions de sens et nouvelles acceptions seront presque aussi nombreuses que les mots neufs. Il en va de même pour les exemples. Certains pourront surprendre par leur simplicité et même leur extrême banalité. Ce n'est pas involontaire. Presque toujours leur présence est destinée à mettre en évidence une construction grammaticale, une règle d'accord, ou l'emploi des prépositions convenables. Usage, usage... Il ne s'agit que d'éclairer le parler de chacun. Pour cette raison, le Dictionnaire, par tradition, ne comporte pas de citations, ni ne fait presque jamais référence nominale à des auteurs. Par discrétion aussi ; les citations, s'il y en avait, seraient par la force des choses empruntées, pour un grand nombre, à des membres disparus ou présents de la Compagnie. »

Le *Trésor de la langue française* (TLF), quant à lui, est le dictionnaire de langue de référence réalisé entre le début des années 60 et le milieu des années 90 par l'Institut National de la Langue Française (INaLF, laboratoire du CNRS) dont le laboratoire ATILF est aujourd'hui le successeur nancéien.

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 et présentés dans la préface du TLFi (ATILF 2004). Le Trésor sera donc :

a) 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.

b) 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.

c) 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.

d) 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. Ainsi, dès 1964 et grâce à un puissant matériel informatique[10], des dépouillements systématiques et exhaustifs de plus de 1 000 œuvres littéraires permirent aux rédacteurs de s'appuyer sur une collection très riche d'exemples d'usage des mots (430 000 exemples). Cette base de données textuelles, initiée pour les besoins du TLF, donna naissance à une des plus grandes bases de données textuelles sur une langue donnée, FRANTEXT, qui, mise à jour régulièrement, regroupe aujourd'hui, à l'ATILF, plus de 4 000 œuvres littéraires françaises.

Le *Trésor de la langue française* est donc 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 et dont le but premier était de fournir des données organisées aux rédacteurs du dictionnaire. 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 gauches et droits dans un document consacré aux cooccurrences, ou encore ordre défini selon les constructions syntaxiques propres à chaque partie du discours dans des documents mis au point,

[10] GAMMA 60 au début, CII 10 070 ensuite, puis Multics : les besoins en informatique de ce vaste projet furent ainsi très intimement mêlés à l'évolution de l'IUCA de Nancy (Institut universitaire de calcul automatique), ancêtre du CIRIL d'aujourd'hui (Centre interuniversitaire de ressource informatique de lorraine).

pour tenter de dominer la masse des attestations des mots les plus fréquents. Ces concordances étaient utilisées pour un premier tri de la documentation et permettaient d'obtenir des contextes élargis parmi lesquels seraient sélectionnés les exemples finalement retenus dans le dictionnaire.

Dans son ouvrage sur les dictionnaires de la langue française, Jean Pruvost (2002) présente ainsi cet ouvrage : *« Ce projet, qui correspond à une entreprise publique ayant requis une centaine de chercheurs pendant 30 ans, avec un dépouillement de plus de 3 000 textes littéraires, scientifiques et techniques, a bénéficié des compétences nationales et internationales les plus éminentes […] Il en résulte, au-delà de la très grande qualité scientifique des articles, une description du fonctionnement de la langue qui ne manque pas d'être impressionnante : 23 000 pages, 100 000 mots, 450 000 entrées, 500 000 citations précisément identifiées. Le TLF relève pleinement d'une lexicographie philologique et historique, recourant aux citations-attestations qui permettent de fonder toutes les analyses morphologiques et sémantiques ».*

1.2. *S*tructure et usage des dictionnaires papier

Chacun de nous connaît la structure de tels dictionnaires papier, nous ne nous appesantirons donc pas trop sur ces structures. Elles se caractérisent par une liste d'entrées rangées par ordre alphabétique stricte donnant accès à chaque article : chaque entrée correspondant à une forme normalisée d'un mot (singulier pour les noms, masculin singulier pour les adjectifs, infinitif pour les verbes). Quant à la structure des articles, elle n'est en fait perceptible qu'au travers de la présentation typographique (cf. figure1) et, si elle ne pose que peu de problèmes à l'humain pour son interprétation, elle demeure souvent rétive à un traitement automatique. En particulier très souvent se trouvent cachées dans des articles des définitions d'autres mots tels `Rhombe` comme synonyme vieilli de losange ou des expressions telles `En losange` toujours dans le même article.

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. : Extrait du TLF papier.

Cette structure induit pour une bonne part les limites ou difficultés d'usage de ces dictionnaires que nous allons expliciter maintenant.

1.2.1. *Limite des approches sémasiologique*

L'approche sémasiologique, du signe au concept, qui structure les articles des dictionnaires à partir des entrées ou vedettes classées par ordre alphabétique, continue de dominer largement la structuration des dictionnaires et impose de fait de connaître le mot recherché pour le trouver. En d'autres termes, cela revient à dire que nos dictionnaires papier classiques sont plus adaptés à la recherche d'un sens ou à sa vérification qu'à la recherche du mot juste pour exprimer une notion ou un concept qui correspond en fait à l'approche inverse ou onomasiologique. Ainsi si vous recherchez dans un dictionnaire le mot correspondant à *un genre d'insectes orthoptères marcheurs, au corps allongé et frêle, dont les espèces sont surtout exotiques, remarquables par leur mimétisme avec les branches, brindilles ou tiges sur lesquelles ils séjournent*, vous n'avez aucune chance de la trouver facilement si vous ne vous rappelez pas qu'il s'agit de `phasmes` !

Cette même approche rend très difficile, pour ne pas dire inopérante, des recherches par champs lexicaux ou par tout autre procédé que les recherches via l'entrée ou la vedette.

1.2.2. *Limite des accès par entrée ou vedette*

L'accès alphabétique par entrée du dictionnaire ou nom de vedette nécessite de connaître à priori la forme canonique d'un mot et son orthographe précise. Si on reprend l'exemple précédent, il convient de plus de savoir que ce sont des `phasmes` et non des `fasmes`. Et dans le cas de verbes à conjugaison irrégulière si on ne connaît pas l'infinitif d'une forme irrégulière on ne pourra accéder à son sens. Pensons à des locuteurs étrangers qui rechercherait la définition de `irons`, s'ils ne savent pas que c'est la forme de la première personne du pluriel du futur du verbe `aller`, ils n'auront que peu de chance de trouver la bonne définition. Ce type de limitation est, hélas, encore trop souvent sous-estimé alors même que plusieurs études ont montré qu'un des premiers usages des dictionnaires est précisément de rechercher ou vérifier l'orthographe d'un mot !

Une autre limite liée à ces accès par entrée ou vedette est due aux nombreuses entrées cachées dans un dictionnaire qui correspondent aux mots ou locutions définies sous une autre entrée. Certes très souvent ce ne sont que des dérivés que l'on va assez facilement retrouver par un accès orthographique. C'est par exemple le cas dans le TLF d'`Écobueur` défini sous `Écubuer` (cf. figure 2).

ÉCOBUER, verbe trans.
AGRIC. Défricher par l'écobuage*. *Les chaumes, écobués par larges places, se recouvraient de cendres grises* (LA VARENDE, *Contes fervents,* Pinsonnière, 1948, p. 96). *Le beau-père avait (...) arraché les souches, écobué les buissons, aplani les cendres et fait une terre (...) la meilleure de toutes pour le blé* (GIONO, *Joie demeure,* 1935, p. 445).

DÉR. Écobueur, subst. masc. Celui qui pratique l'écobuage. *Les incendies doivent être attribués à l'imprudence des chasseurs, des passants et des écobueurs* (*Enquêtes sur les incendies de forêts,* p. 77 ds LITTRÉ). Attesté par GUÉRIN 1892 et par *Lar. 19e Suppl.* 1878-*Lar. encyclop.* Seule transcr. ds LITTRÉ : é-ko-bu-eur.

Figure 2. : Un exemple simple d'entrée cachée pour un dérivé.

Mais le *Trésor de la langue française* qui possède un nombre non négligeable d'entrées correspondant à des éléments formants tels *-PATHIE, -PATHIQUE, -PATHE* rend difficilement accessibles les nombreux mots définis sous cette entrée dont certains très utilisés aujourd'hui, telle `Empathie`, auraient peut-être mérité d'une entrée spécifique (cf. figure 3.).

Ce phénomène ne se limite pas aux seuls dérivés ou éléments formants, on peut retrouver dans un dictionnaire des synonymes et antonymes dont la définition se trouve cachée sous une autre entrée.

Comme nous allons le voir dans la suite une bonne informatisation d'un dictionnaire va permettre de dépasser ces limites des dictionnaires papier et permettre ainsi d'ouvrir vers des usages nouveaux des dictionnaires.

1.3. Du dictionnaire papier au dictionnaire informatisé : l'exemple du TLFi

Reflet fidèle de la version papier, le TLFi[11] se caractérise, comme le TLF, par la richesse de son matériau et la complexité de sa structure :

[11] www.atilf.fr/tlfi

a) Importance de sa nomenclature : 100 000 mots avec leur étymologie et leur histoire et 270 000 définitions.

b) Richesse de chaque article (vedettes, codes grammaticaux, indicateurs sémantiques ou stylistiques, indicateurs de domaines, définitions, exemples référencés…).

c) Richesse des 430 000 exemples, tirés de plus de deux siècles de production française.

d) 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 bibliographique.

***PATHIE**, -PATHIQUE, -PATHE, élém. Formants*

Élém. tirés du gr. «ce que l'on éprouve [de mal]» et entrant dans la constr. de mots sav. appartenant notamment au vocab. de la méd.; *-pathie* est empr. au gr. ; les subst. fém. constr. peuvent générer des adj. dér. en *-pathique* et des dér. régressifs en *-pathe,* adj. ou subst. masc.; à noter l'adj. *protopathique* (*infra* I A) qui semble usité sans terme en *-pathie* correspondant.

I. [*-pathie* exprime une manière d'être affecté, de sentir, de ressentir, caractérisée par le 1er élém.; celui-ci est le plus souvent issu du gr.]

A. [Les mots constr. désignent ou expriment un rapport avec des phénomènes sensitifs d'ordre pathol.] :

*hyper***pathie.** Sensibilité exagérée à la douleur, qui s'observe notamment dans les syndromes thalamiques du côté de la lésion cérébrale. La douleur est très pénible et persiste après l'arrêt de l'excitation (*Méd. Biol.* t.2 1971). *Hyper***pathique.** Qui se rapporte à l'hyperpathie *(Méd. Biol.* t.2 1971).

*proto***pathique,** psychol. [P. oppos. à *épicritique (s.v. épi-)*] Qui est apte à percevoir uniquement des stimulations sensitives, tactiles ou thermiques, grossières (*Méd. Biol.* t.3 1972). *Même dans les expériences de Head et Rivers, où les paliers sont nets, au cours de la régénération du nerf, la sensibilité épicritique reparue ne supprime pas entièrement la sensibilité protopathique* (RUYER, *Conscience,* 1937, p.83). […]

B. [Les mots constr. désignent un rapport à autrui]:

***intro*pathie** (*intro-,* v. *intra* rem. gén. 2). Synon. de ***empathie** (infra)* (d'apr. MARCH. 1970). *La connaissance de moi-même est toujours à quelque degré un guide dans le déchiffrement d'autrui, bien qu'autrui soit d'abord et principalement une révélation originale de l'intropathie* (RICOEUR, *Philos. volonté,* 1949, p.14).

Rem. 1. Empr. **a)** Au lat. du gr., v. *antipathie, apathie, sympathie.* **b)** À l'angl., v. *télépathie.* **2.** Adapt. du gr. **Empathie,** psychol. Subst. fém., Mode de connaissance par une forme de sympathie qui atteint autrui en lui-même, sans toutefois s'identifier à lui (FOULQ. 1971). **Empathique**, adj., Qui relève de l'empathie. *L'identification du bébé et de la maman, sur le mode empathique correspondant par exemple à l'incorporation orale de la mère* (*Traité sociol.,* 1968, p.409). *L'empathie ou la perception empathique (...) c'est de percevoir le monde subjectif d'autrui «comme si on était cette personne sans toutefois perdre de vue qu'il s'agit d'une situation comme si...»* (C. ROGERS, M. KINGET, *Psychothérapie...,* I, 1963, p.187-188 ds FOULQ. 1971).

Figure 3. : Un exemple plus complexe d'entrée cachée sous la définition de l'élément formant *-PATHIE* dans le *TLF.*

Cette version du TLFi (Dendien & 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.

1.3.1. *Saisie initiale du dictionnaire*

La première étape d'informatisation du TLF a consisté à réaliser une archive fiable de la totalité des seize tomes papier sur support informatique. 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. 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. 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. 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.

1.3.2. *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[12].

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 a donc consisté à enrichir le texte initial d'annotation XML de forme, de fond et de structure. 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.

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

a) 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.

```
<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>En losange.</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>
```

Figure 4. : Balisage typographique.

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

b) *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 bâtie 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 5).

```
<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>1878-1935</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 5. : Enrichissement par balisage de contenu.

c) *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 illustrées par 427 493 exemples,
- nombre total de balises XML : 36 613 712,
- niveau de profondeur hiérarchique maximal : 23.

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

(i) Base lexicale pour une hypernavigation

Tous les mots d'une page Internet peuvent être considérés comme autant d'hyperliens virtuels. L'effet d'un double clic sur un mot quelconque d'un article peut ainsi 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) et la cible (celle qui traite le mot associé) d'une hypernavigation. Lors de la sélection d'un mot, nous ne transmettons que la forme à l'application cible. C'est donc elle qui procède éventuellement à la lemmatisation de la forme (en absence de contexte, plusieurs lemmes sont parfois envisageables). En cas de sélection d'un mot, 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 R) ; 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 historique du vocabulaire français* (BHVF : dictionnaire des datations de l'origine des mots réalisé par notre laboratoire) ; enfin, la base de données FRANTEXT. On s'imagine 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 (BHVF : dictionnaire des datations de l'origine des mots réalisés par notre laboratoire).

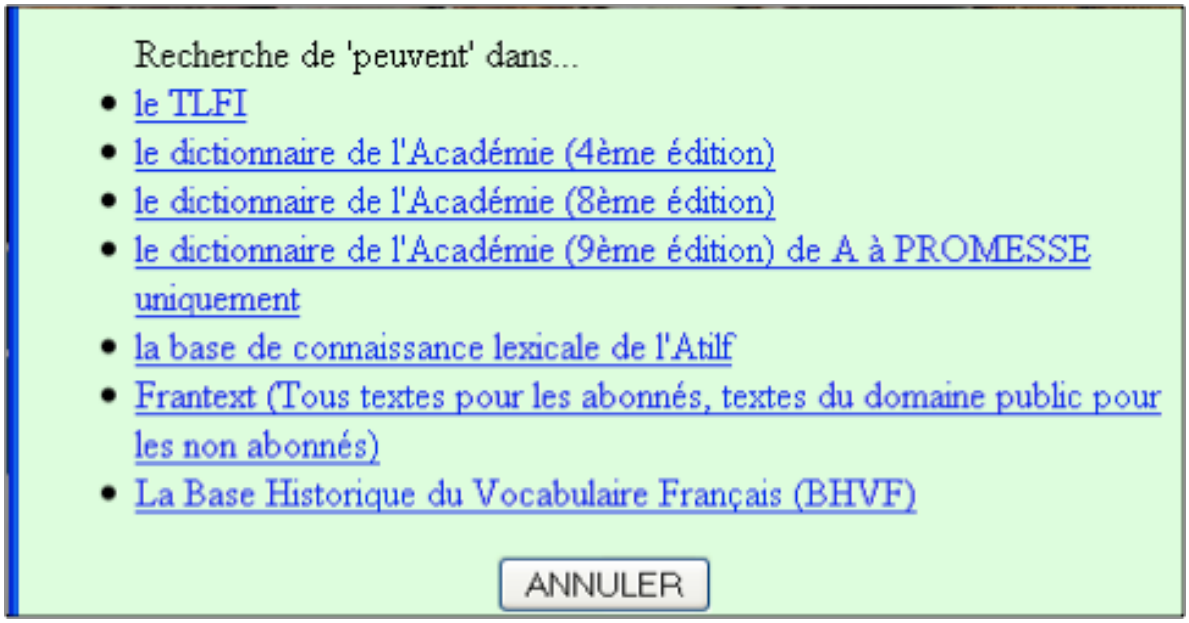
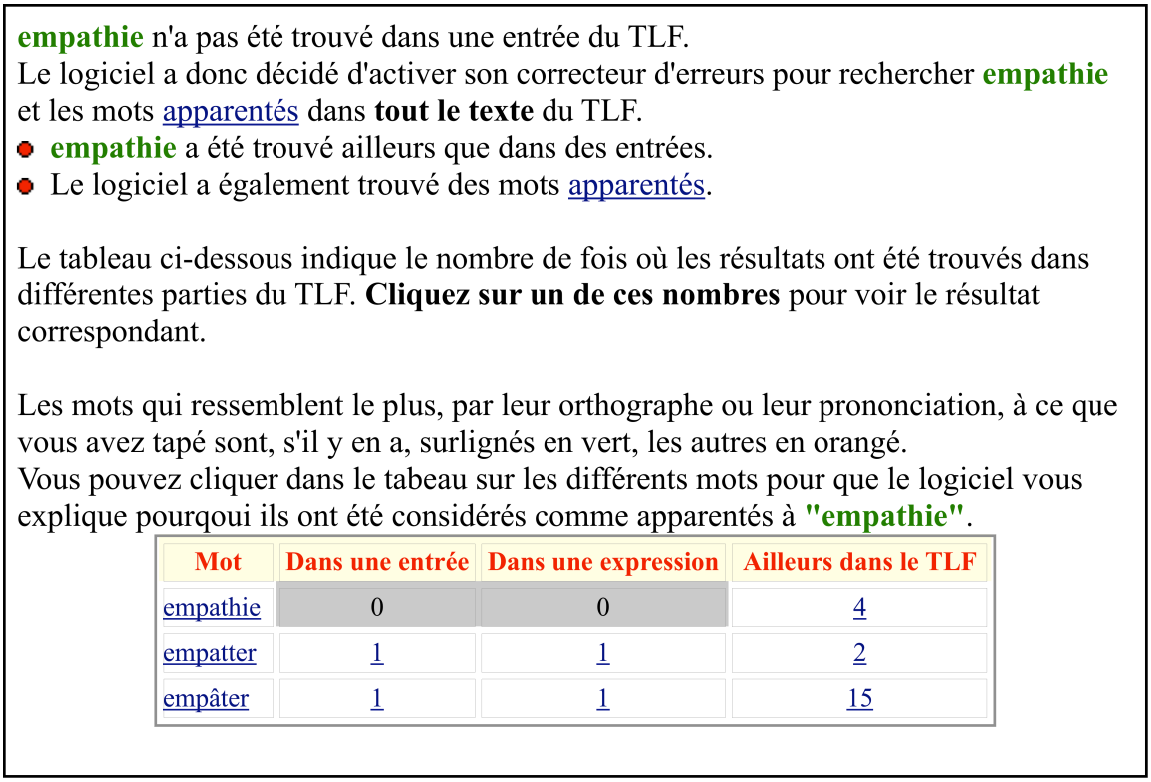


Figure 6. : Fenêtre d'hypernavigation dans le TLFi.

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). L'ensemble des informations de cette base (lemmes, formes, informations associées) est codé en XML, 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[13]. 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. À l'image 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 formes du verbe `savoir`. L'exploitation d'une telle base de données lexicales permet ainsi de dépasser certaines limites d'accès aux dictionnaires que nous avons signalées ci-dessus. Elle permet en particulier de pouvoir interroger le dictionnaire à partir d'une forme quelconque et non plus uniquement d'un lemme.

Notons de plus que l'ensemble du dictionnaire est ouvert à une recherche plein texte, ce qui permet de retrouver sans difficulté les entrées cachées. Ainsi si l'on reprend le cas d'`empathie`, une requête sur ce mot sur la version web ou la version cédérom du TLFi permet d'obtenir le résultat suivant (Figure 7) :



empathie n'a pas été trouvé dans une entrée du TLF.
Le logiciel a donc décidé d'activer son correcteur d'erreurs pour rechercher **empathie** et les mots apparentés dans **tout le texte** du TLF.

- **empathie** a été trouvé ailleurs que dans des entrées.
- Le logiciel a également trouvé des mots apparentés.

Le tableau ci-dessous indique le nombre de fois où les résultats ont été trouvés dans différentes parties du TLF. **Cliquez sur un de ces nombres** pour voir le résultat correspondant.

Les mots qui ressemblent le plus, par leur orthographe ou leur prononciation, à ce que vous avez tapé sont, s'il y en a, surlignés en vert, les autres en orangé.
Vous pouvez cliquer dans le tabeau sur les différents mots pour que le logiciel vous explique pourqoui ils ont été considérés comme apparentés à **"empathie"**.

Mot	Dans une entrée	Dans une expression	Ailleurs dans le TLF
empathie	0	0	4
empatter	1	1	2
empâter	1	1	15

Figure 7. : Exemple de résultats de recherche d'un mot autre qu'une entrée du dictionnaire.

(ii) Base de données phonétiques pour un accès tolérant aux fautes d'orthographe

Comme nous l'avons présenté dans (Dendien & Pierrel 2003) 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

[13] accessible à l'adresse www.cnrtl.fr/lexiques/morphalou

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.

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.

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 déplaît 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 :

a) absence systématique d'accents ou de cédilles (pratique liée au contexte informatique ?),

b) 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),

c) 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,

d) 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 importante. 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.

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

b) La prise en compte du fait que l'utilisateur a peut-être fourni une forme flexionnelle au lieu d'un lemme : il est alors nécessaire de procéder aux lemmatisations nécessaires. L'assistance de la base de données lexicales que nous avons décrite plus haut permet de traiter cet aspect.

c) L'acceptation que la recherche de l'utilisateur peut porter sur les entrées du dictionnaire bien entendu, mais aussi sur tout le texte du dictionnaire. 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.

d) La nécessité de permettre à l'utilisateur de fournir un équivalent phonétique de ce qu'il cherche. Cela nécessite la création de ressources permettant des traitements phonétiques pour 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. C'est ce que nous avons mis en œuvre dans le TLFi.

La combinaison de ces divers traitements qui sont de trois ordres (correction orthographique, traitement phonétique, lemmatisation) nous a permis de développer un accès très tolérant aux fautes des utilisateurs.

Reprenons à titre d'illustration le cas d'un utilisateur ayant tapé « `jenero` » (Dendien & Pierrel 2003). 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 de phonétisation 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 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 pas s'ils doivent être écrits attachés ou non (*anticapitalisme*, *paranormal*, etc.). Il est amusant de constater qu'un tel mécanisme dote également le logiciel du TLF de la possibilité d'appréhender les calembours (rechercher « `aile et faon` » ou « `sauces y sont` » !).

2. Dictionnaires informatisés et usages nouveaux

2.1. L'exemple du TLFi

On peut trouver à l'adresse www.tlfi.fr une présentation et des démonstrations sur les recherches offertes 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.

2.1.1. *Recherche d'un mot ou d'une expression*

Cette recherche permet un accès à un mot à travers un système de correction automatique (forcée ou non) : ainsi, en introduisant la recherche de la forme `etique` (sans accent), on accède aux deux articles correspondants 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'expression telle `le trompette` en focalisant la réponse sur l'élément pertinent demandé et en offrant la possibilité de surligner tel ou tel objet textuel, ici par exemple les définitions :

> Objets de la recherche : Paragraphe
> TROMPETTE, subst.
>
> II. *Subst. masc.* Personne qui joue de la trompette.
> A. 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).
> *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).

2.1.2. *Recherche assistée*

Ce second type d'accès permet par exemple de rechercher des expressions composées d'une forme : ainsi, en demandant les mots contenant la forme *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.
Etc.

ou 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, dans le domaine de la marine, correspondent à une définition incluant une forme du mot *voile*, soit dans une structure plus compacte :

[code grammatical : *verbe* ; domaine : *marine* ; type d'objet : *définition*, contenu : &m*voile*[14]]. Voici un extrait des 61 réponses que l'on obtient :

ABRIER, ABREYER, verbe trans.
Empêcher le vent, en l'interceptant, de passer jusqu'à (une autre **voile**) :
AGRÉER[2], verbe trans.
„Préparer ou travailler à la garniture, aux agrès d'un bâtiment, fourrer les dormans, estroper les poulies, garnir **voiles**, vergues, etc. : `` (WILL. 1831) :
AMURER, verbe.
Fixer l'amure d'une **voile** pour l'orienter selon le vent :
Etc.

ou encore l'ensemble des mots dont la définition utilise le mot `liberté` [type d'objet : définition, contenu : &mliberté] ; on obtient ainsi 306 réponses dont :

ABUSER, verbe trans.

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

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

2 (Celui) à qui on a donné la **liberté**.

AISE[1], subst. fém.

3 Grande **liberté**.

4 **Liberté** et souplesse totale des mouvements du corps :

5 **Liberté** et facilité.

ALIÉNANT, ANTE, part. prés. et adj.

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

ALIÉNATION, subst. fém.

7 Privation de **libertés**, de droits humains essentiels éprouvée par une personne ou un groupe social sous la pression de facteurs permanents (Hegel) ou historiques (Marx) qui l'asservissent à la nature ou à une classe dominante.

ÂME, subst. fém.

8 Aliéner sa **liberté**, sa dignité... en échange de quelque chose : **(Vendre son âme (au diable).)**

AMPLIER, verbe trans.

9 Lui accorder plus d'aisance, lui donner plus de **liberté** dans la prison. **(Amplier un prisonnier (Ac. Compl. 1842, LITTRÉ et Lar. 19e).)**

2.1.3. *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 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*.

[14] &m*subs* permet de tester toutes les formes d'un *substantif,* de même que &c*verbe* toutes les formes d'un *verbe.*

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 :

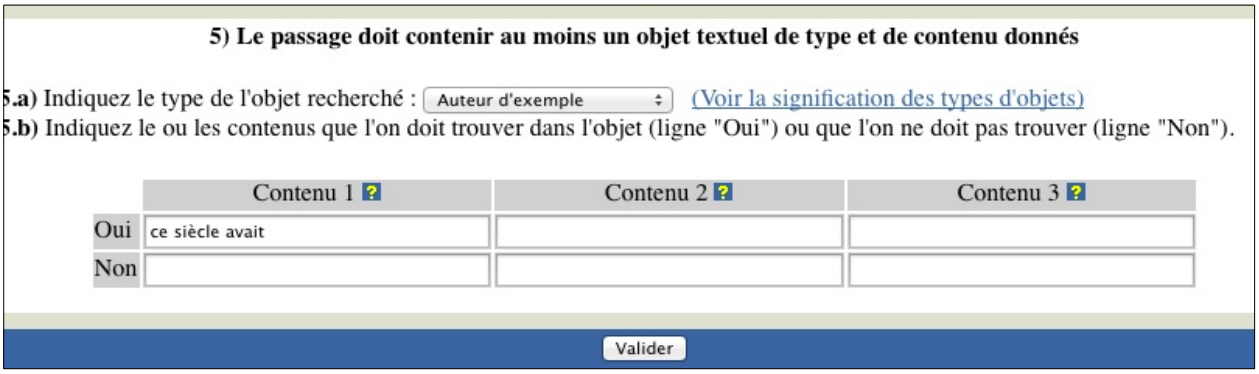
Objets de la recherche : *Entrée ; Code grammatical ; Domaine technique ; Langue empruntée*

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

2.1.4. *Recherche de citations*

Un dernier usage souvent exploité dans le TLF est la recherche de citations. En effet, le TLF contient environ 430 000 exemples d'usage avec leur référence. Ces exemples ayant été soigneusement 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 correspondants à sa demande soit sous forme simplifiée (cf. Figure 8) soit sous forme étendue (cf. Figure 9) et retrouve ainsi la citation complète avec ses références précises.

Objets de la recherche : 1Exemple1

PERCER, verbe
1 1**Ce siècle avait** deux ans! Rome remplaçait Sparte, Déjà Napoléon perçait sous Bonaparte, Et du Premier Consul, déjà, par maint endroit, Le front de l'Empereur brisait le masque étroit (HUGO, Feuilles automne, 1831, p. 717).1
SIÈCLE, subst. masc.
2 1**Ce siècle avait** deux ans! Rome remplaçait Sparte, Déjà Napoléon perçait sous Bonaparte (HUGO, Feuilles automne, 1831, p. 717).1
SOUS, prép.
3 1**Ce siècle avait** deux ans! Rome remplaçait Sparte, Déjà Napoléon perçait sous Bonaparte, Et du premier Consul, déjà, par maint endroit, Le front de l'Empereur brisait le masque étroit (HUGO, Feuilles automne, 1831, p. 717).1

Figure 8. : Résultats simplifiés

Objets de la recherche : 1Exemple1

H **PERCER, verbe**

C. — *P. anal.* Apparaître, se montrer. *Le soleil perce. Le jour perce à peine à travers les vitraux* (STAËL, *Corinne,* t. 2, 1807, p. 142).
D. — *Au fig.*
1. Se manifester. *La bonne humeur du Roi, depuis que la révolte contre le bailli lui avait été annoncée, perçait dans tout* (HUGO, *N.-D. Paris,* 1832, p. 507). *Son dédain pour la philosophie perçait à chaque mot; c'était un perpétuel sarcasme* (RENAN, *Souv. enf.,* 1883, p. 235).
◆ *P. allus. littér.* 1***Ce siècle avait*** *deux ans! Rome remplaçait Sparte, Déjà Napoléon perçait sous Bonaparte, Et du Premier Consul, déjà, par maint endroit, Le front de l'Empereur brisait le masque étroit* (HUGO, *Feuilles automne,* 1831, p. 717).1

Figure 9. : Résultat étendu.

2.1.5. *L'impact du TLFi*

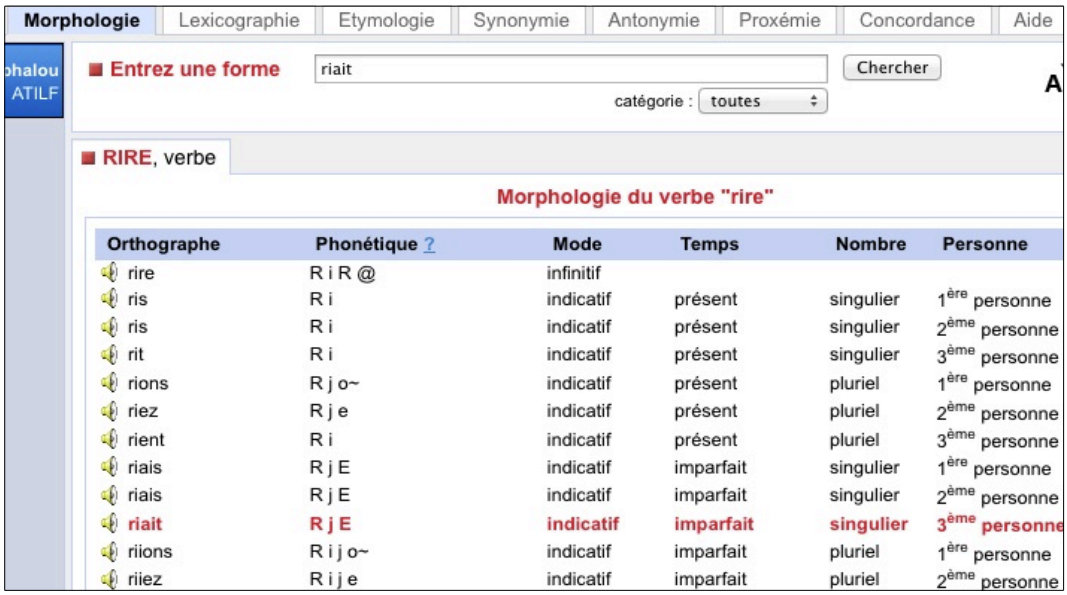
Le TLFi, sans aucun doute le plus grand dictionnaire informatisé consacré à la langue française, grâce à la richesse de son contenu entièrement encodé en XML, a ouvert des perspectives intéressantes. Sa mise à disposition sous forme de Cédérom et sur le Web a rencontré un succès important tant auprès du grand public que des utilisateurs universitaires ou des professionnels de la langue : objet de plusieurs centaines de milliers de connexions quotidiennes en provenance de tous les continents, il est référencé par d'innombrables sources et la notoriété qu'il a acquise en fait un outil de promotion appréciable de la langue française. Alors même que le TLF avait la réputation tenace d'être un dictionnaire réservé à une élite, sa version informatique et les interconnexions par hypernavigation avec le dictionnaire de l'Académie, FRANTEXT ou la base historique du vocabulaire français le positionnent au cœur d'un ensemble de ressources sur la langue française. Ses usages actuels, plusieurs centaines de milliers d'accès par jour prouvent qu'il joue un rôle actif et prépondérant dans la valorisation de notre langue, démontrant ainsi que sa réputation élitiste est injustifiée.

2.2. Un exemple d'intégration de ressources : le portail lexical du CNRTL[15]

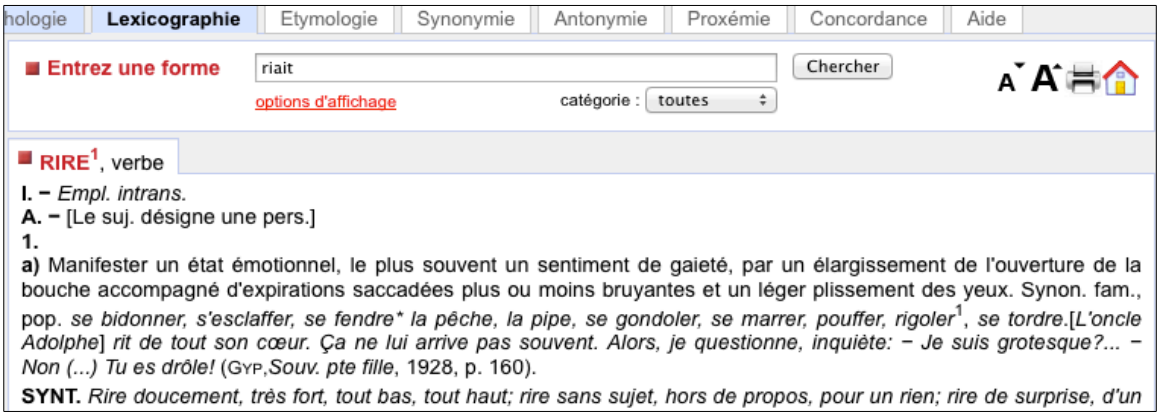
Le portail lexical du CNRTL (Pierrel & Petitjean 2007) a pour vocation de valoriser et de partager, en priorité avec la communauté scientifique, un ensemble de données issues des travaux de recherche sur le lexique français. Projet évolutif, cette base de connaissances lexicales exploite aujourd'hui divers documents numériques pour fournir, à partir d'une forme lexicale, cinq types d'informations importantes : des informations morphologiques issues de MORPHALOU, des informations lexicographiques et étymologiques issues des projets TLF et TLF-Etym[16] des informations de synonymies à travers l'intégration du *Dictionnaire de synonymes de Caen*[17] et une concordance utilisant le corpus des textes de la base FRANTEXT. Il offre aussi la possibilité d'exporter les résultats du concordancier au format XML/TEI. C'est à notre connaissance le seul site permettant à un utilisateur d'exporter dans un format normalisé un concordancier français d'une telle importance.

Voici à titre d'exemple le type de résultat accessible via ce portail :

a) Informations morphologiques, accessibles directement pour la forme *riait* par la requête http://www.cnrtl.fr/morphologie/*riait*

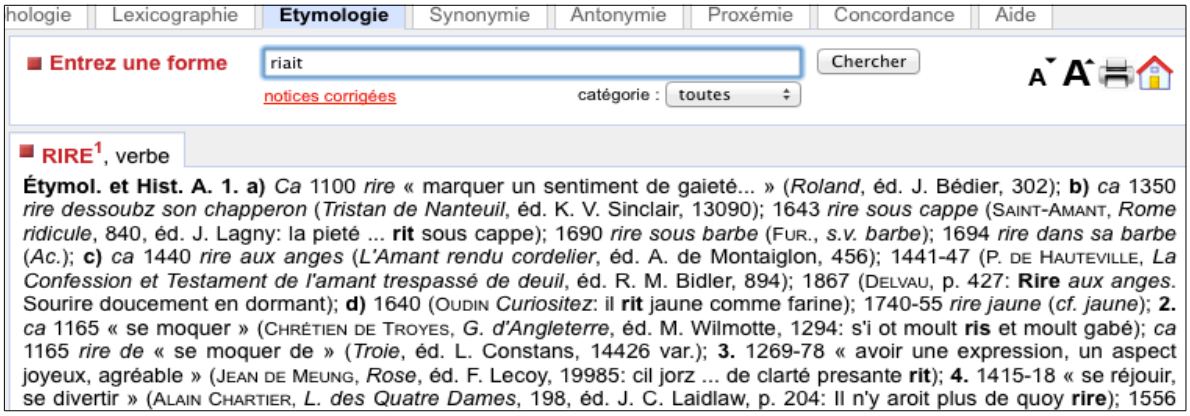


b) Informations lexicographiques, accessibles directement pour la même forme par la requête http://www.cnrtl.fr/lexicographie/*riait*

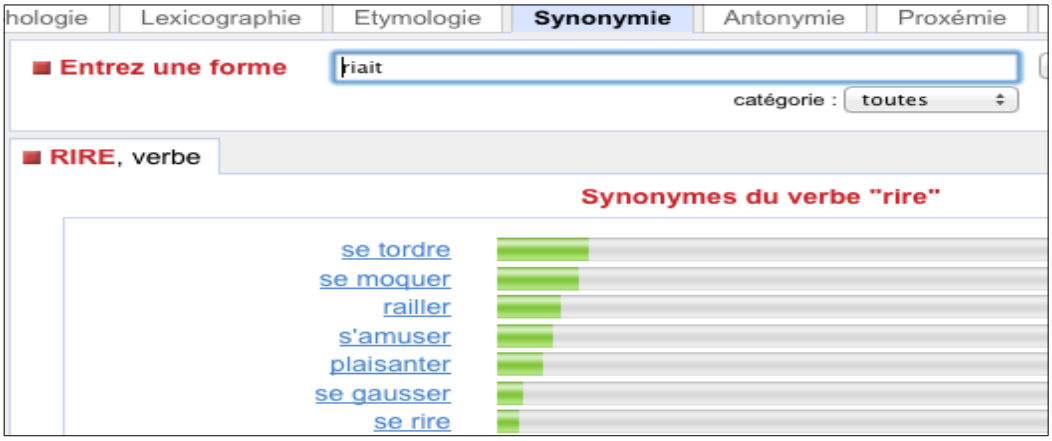


[15] www.cnrtl.fr/portail
[16] www.atilf.fr/tlf-etym
[17] http://www.crisco.unicaen.fr/

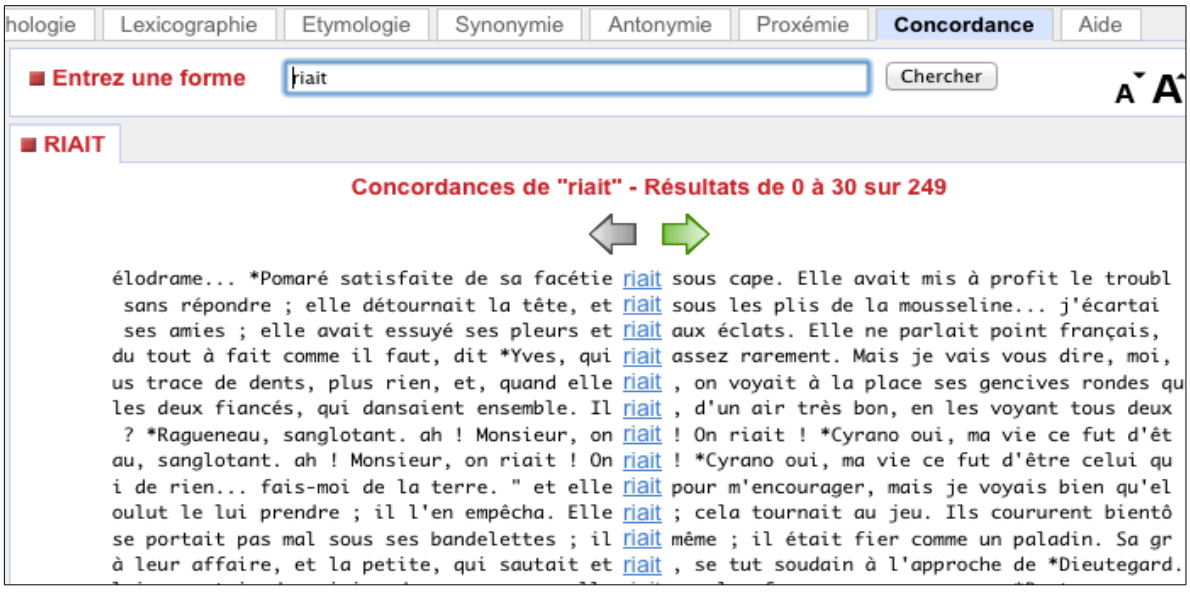
c) Informations étymologiques, accessibles directement par la requête http://www.cnrtl.fr/etymologie/*riait*



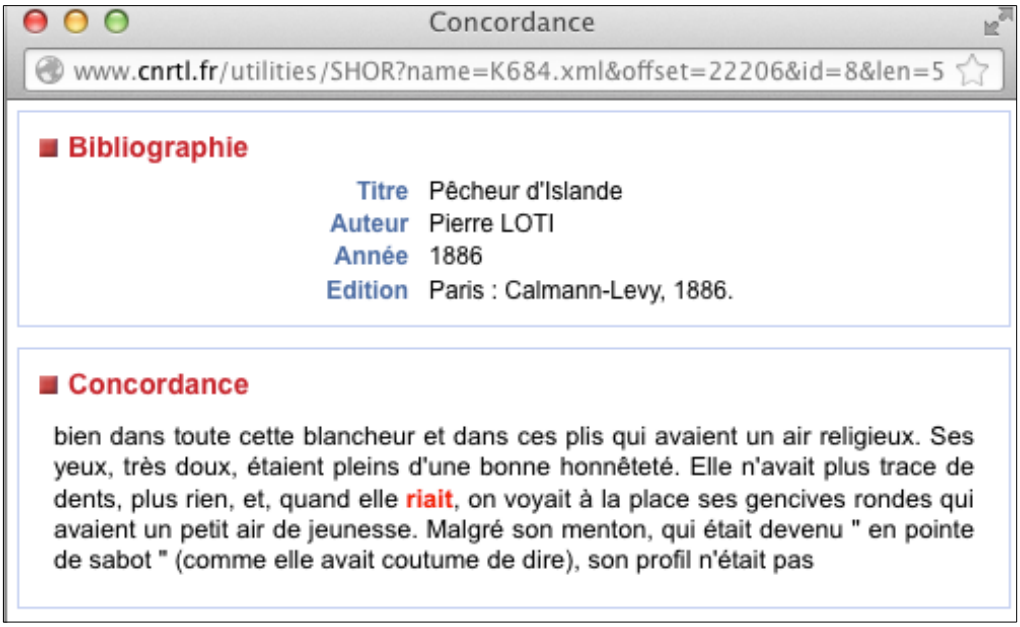
d) Informations de synonymie, accessibles directement par la requête http://www.cnrtl.fr/synonymie/*riait*



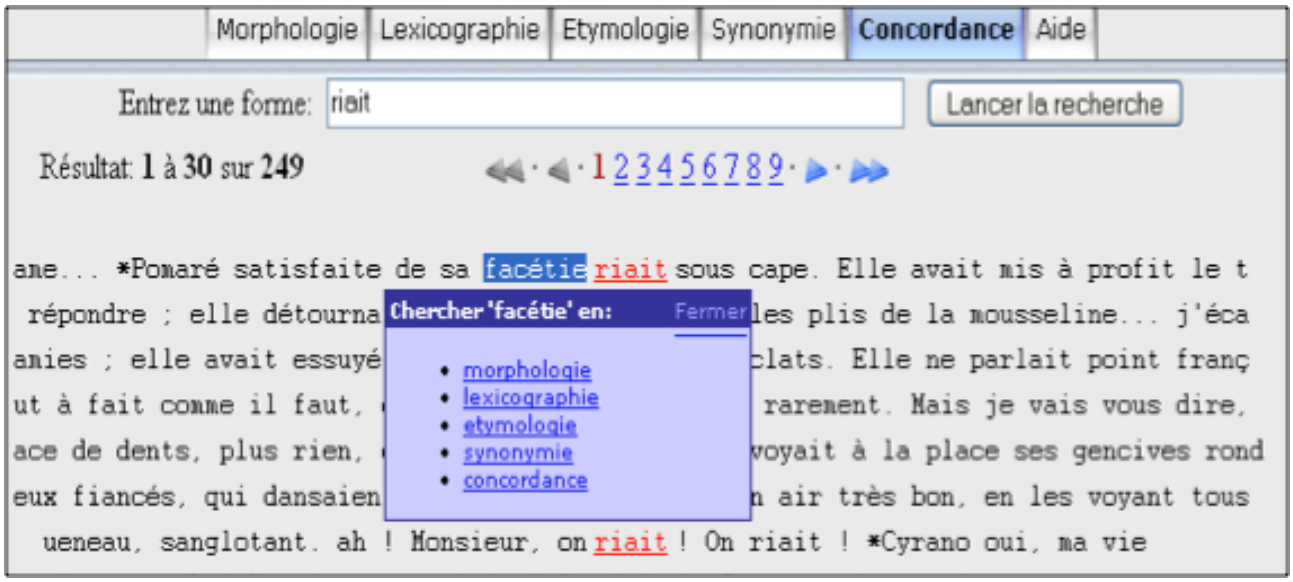
e) concordances, accessibles directement par la requête http://www.cnrtl.fr/concordance/*riait*



f) De plus, un simple clic droit sur un des exemples permet d'obtenir la référence complète de l'exemple sélectionné, ainsi pour le premier exemple :



Le portail lexical permet également, à partir d'un simple double-clic sur un mot, une hypernavigation vers toutes les informations lexicales disponibles pour ce mot. Par exemple, si l'on veut obtenir des informations sur le mot « facétie » du premier exemple de concordance, un double-clic sur le mot affiche un menu qui permet d'hypernaviguer vers les informations lexicales de ce mot :



3. Vers une véritable ressource lexicale pour le TAL

On peut répartir en trois types les ressources lexicales informatisées sur la langue française actuellement disponibles pour une exploitation en traitement automatique des langues :

a) des ressources de type dictionnaire informatisé, à l'image du TLFi, celles-ci sont avant tout utiles pour un usage humain et elles ne sont pas suffisamment structurées (en données élémentaires stockées sous forme de base de données, par exemple) et formalisées pour un usage en traitement automatique ;

b) des bases de données lexicales, certaines contenant des informations très ciblées sur les mots, par exemple MORPHALOU pour la morphologie (cf. supra), d'autres plus riches, telles les ressources développées dans le cadre du LADL[18] ou la base de verbes Dicovalence développée par Piet Mertens (Mertens 2010)[19] ;

c) des réseaux de type *WordNet* (Fellbaum 1998), dont les fondements linguistiques sont discutables (Slodzian 2000). De plus, pour le français, le réseau lexical *EuroWordNet* (De Loupy, Dutoit, El-Bèze & Griot 1999), n'a qu'une couverture très partielle et reste difficilement accessible pour une exploitation commerciale en raison de problèmes de licence.

Il existe aussi des projets visant la construction automatique de réseaux ou, plus généralement, de ressources lexicales à partir de la compilation et de la fusion de données présentes dans des ressources existantes, par exemple la ressource WOLF (Sagot & Fišer 2008). Le principal problème de telles approches est qu'elles favorisent la couverture au détriment de la précision et de l'exactitude de l'information encodée. Elles permettent donc de générer des ressources potentiellement très utiles pour certains gros traitements statistiques, mais tout à fait inappropriées pour de la recherche sémantique fine, notamment en raison de l'absence d'une bonne gestion de la polysémie. De plus, de telles ressources ne peuvent en aucun cas être utilisées comme des descriptions de référence sur le français, dans un contexte de consultation humaine ou un contexte pédagogique. Enfin, ces approches présupposent l'existence de ressources en accès libre de bonne qualité. Or ces dernières font largement défaut, notamment pour le français.

À titre d'exemple, il n'y a pas WOLF d'entrée proprement dite pour un mot tel `abdiquer`. Chaque entrée décrit en fait, à l'image de Wordnet, un « synset » (c'est-à-dire, un ensemble de quasi-synonymes) et la modélisation du vocable en question est enchâssée dans l'entrée d'un synset (ou de plusieurs, en cas de polysémie effectivement décrite). Dans l'exemple de la figure 10, seul le sens premier de `abdiquer` (acte officiel) est présent dans WOLF, enchâssé dans un synset qui se trouve ne contenir qu'un élément (un seul « littéral » entre les balises <SYNONYM> et </SYNONYM>). On constate que non seulement l'information donnée est absolument minimale, mais qu'en plus la polysémie du vocable n'est pas prise en compte (pas de prise en compte du sens présent dans des expressions comme *abdiquer ses responsabilités*). On peut contraster ce type de description avec les informations données dans une ressource pédagogique telle que le *Lexique actif du français* (Mel'čuk & Polguère 2007), qui est construite à partir d'une base lexicale formelle du français et dont on présente la même entrée en figure 11.

```
<SYNSET>
<ID>ENG20-02308091-v</ID>
<POS>v</POS>
<SYNONYM>
<LITERAL>abdiquer<SENSE>0/2:enwiktionary,frwiktionary</SENSE></LITERAL>
</SYNONYM>
<ILR><TYPE>hypernym</TYPE>ENG20-02296272-v</ILR>
<ILR><TYPE>eng_derivative</TYPE>ENG20-00194287-n</ILR>
<ILR><TYPE>eng_derivative</TYPE>ENG20-06109588-n</ILR>
<ILR><TYPE>eng_derivative</TYPE>ENG20-06809685-n</ILR>
<ILR><TYPE>eng_derivative</TYPE>ENG20-06809823-n</ILR>
<ILR><TYPE>eng_derivative</TYPE>ENG20-09136863-n</ILR>
<DEF>give up, such as power, as of monarchs and emperors, or duties and
obligations</DEF>
<USAGE>The King abdicated when he married a divorcee</USAGE>
<DOMAIN>administration</DOMAIN>
<SUMO>Giving<TYPE>+</TYPE></SUMO>
</SYNSET>
```

Figure 10 : Codage XML d'un Synset contenant la description de `abdiquer` dans WOLF.

[18] http://infolingu.univ-mlv.fr/

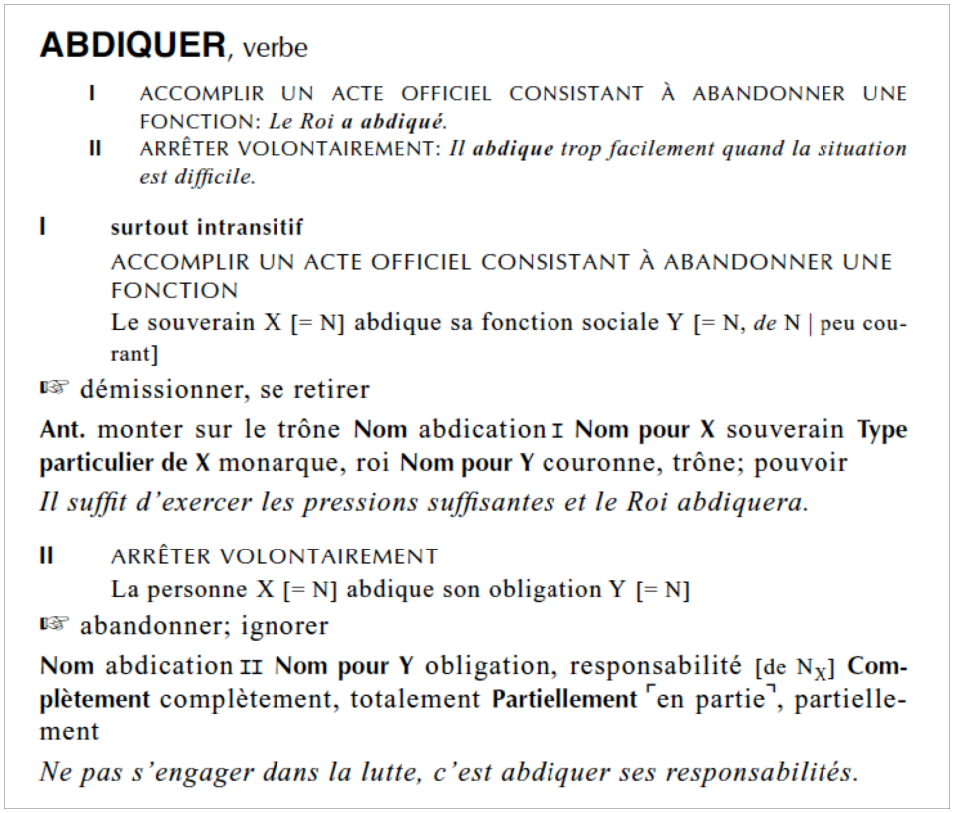
[19] http://bach.arts.kuleuven.be/dicovalence/

Le projet RELIEF que nous développons à Nancy vise la construction d'une ressource lexicale du français à large couverture, le RLF, développé à partir du fonds lexicographique que représente le *Trésor de la langue française informatisé* (TLFi). Il s'agit d'un projet ambitieux[20] – destiné à se poursuivre après cette phase initiale – qui propose une restructuration, une réécriture et une mise à jour des données lexicales actuellement disponibles au sein du TLFi pour rendre compte des néologies de forme et des usages nouveaux (néologies de sens) apparus au cours des dernières décennies. Le RLF se fixe comme objectif d'être à terme une ressource lexicale de référence pour notre langue du fait :

- de sa large couverture,

- de la qualité des données qu'elle contiendra,

- de son exploitabilité dans des contextes applicatifs jugés prioritaires, tels que le traitement automatique du français, la recherche en linguistique française et la didactique du français.



ABDIQUER, verbe

I ACCOMPLIR UN ACTE OFFICIEL CONSISTANT À ABANDONNER UNE FONCTION: *Le Roi a abdiqué.*
II ARRÊTER VOLONTAIREMENT: *Il abdique trop facilement quand la situation est difficile.*

I **surtout intransitif**
ACCOMPLIR UN ACTE OFFICIEL CONSISTANT À ABANDONNER UNE FONCTION
Le souverain X [= N] abdique sa fonction sociale Y [= N, *de* N | peu courant]
☞ démissionner, se retirer
Ant. monter sur le trône **Nom** abdication I **Nom pour X** souverain **Type particulier de X** monarque, roi **Nom pour Y** couronne, trône; pouvoir
Il suffit d'exercer les pressions suffisantes et le Roi abdiquera.

II ARRÊTER VOLONTAIREMENT
La personne X [= N] abdique son obligation Y [= N]
☞ abandonner; ignorer
Nom abdication II **Nom pour Y** obligation, responsabilité [de N_X] **Complètement** complètement, totalement **Partiellement** ⌜en partie⌝, partiellement
Ne pas s'engager dans la lutte, c'est abdiquer ses responsabilités.

Figure 11 : Entrée du vocable abdiquer dans le Lexique actif du français.

Le projet RELIEF s'appuie donc sur les acquis précédents développés à Nancy (TLF, TLFi et Portail lexical du CNRTL) mais l'architecture du RLF (Lux-Pogodalla & Polguère 2011), la nouvelle ressource lexicale développée dans le cadre de ce projet, repose aussi en grande partie sur les recherches menées antérieurement à Montréal en lexicologie et lexicographie par Alain Polguère, maintenant à l'ATILF, et son collègue de l'Université de Montréal Igor Mel'čuk. Ces travaux, effectués dans le cadre des activités de l'Observatoire de linguistique Sens-Texte (OLST)[21], ont déjà conduit à la réalisation de ressources lexicales de type base de données (cf. le *DiCo* et sa version en ligne version *Dicouèbe*[22]) ou dictionnaires (par exemple, le *Lexique actif du français* mentionné ci-dessus). L'approche de la Lexicologie Explicative et Combinatoire (Mel'čuk, Clas & Polguère

[20] Ce projet, soutenu conjointement, pour une première phase de trois ans, par la Région Lorraine au travers de son Agence de Mobilisation Economique et par le FEDER Lorrain, mobilise à l'ATILF une équipe de 10 personnes encadrées par notre collègue Alain Polguère et rémunérées sur ce contrat.
[21] http://olst.ling.umontreal.ca
[22] http://olst.ling.umontreal.ca/dicouebe/

1995), sur laquelle reposent ces travaux, a aussi donné lieu à des applications lexicographiques pour des langues autres que le français (notamment, russe, espagnol et anglais). Cette approche théorique et descriptive de la modélisation des lexiques a fait ses preuves de son applicabilité en modélisation des lexiques. De plus, elle a déjà inspiré des applications de traitement automatique de la langue, en génération automatique de textes notamment. Les travaux de l'OLST ont aussi été une source d'inspiration pour des logiciels commerciaux, telle la suite *Antidote* (de Druide informatique)[23], qui a introduit dans ses dictionnaires, joints au correcteur, un dictionnaire de « cooccurrences », selon un mode d'identification et de présentation inspiré des principes de modélisation des collocations de la *Lexicologie explicative et combinatoire* (Charest, Brunelle, Fontaine & Pelletier 2007].

La structure d'un article du RLF est très similaire à celle de *DiCo* (Jousse & Polguère 2005). Sa conception repose sur l'utilisation d'un éditeur lexicographique (Gader, Lux-Pogodalla & Polguère 2012) d'aide à l'automatisation et la semi-automatisation de certaines tâches lexicographiques (liées notamment à l'ébauche et à la validation des descriptions) et permettant tout à la fois d'engendrer, à partir du réseau lexical ainsi construit, une base de données SQL pour une exploitation en TAL et d'éditer une version lisible de l'article lexicographique correspondant pour un usage par un humain. On trouvera en figure 12 un extrait de l'entrée du vocable `admirer` généré par l'éditeur lexicographique du RLF. Un article du RLF est divisé en six grandes parties :

– caractéristiques grammaticales,

- définition,

- régime grammatical,

- liens lexicaux s'appuyant sur une description en termes de fonctions lexicales,

- exemples extraits de ressources textuelles telles FRANTEXT,

- locutions ou phraséologies spécifiques liées à cette entrée.

Le contenu de la base RLF est mesuré selon les cinq paramètres suivants :

1. le nombre de *vocables* possédant une entrée dans le RLF ;
2. le nombre de *lexies* ayant un article dans le RLF, une lexie étant une unité lexicale, c'est-à-dire une acception spécifique d'un vocable – une lexie doit être vue formellement comme un nœud dans le graphe lexical du RLF ;
3. le taux de polysémie du RLF, c'est-à-dire le nombre moyen d'acceptions par vocable ;
4. le nombre de liens de *fonctions lexicales* (Mel'čuk & Polguère 1995) connectant les lexies du RLF entre elles – un lien de fonction lexicale doit être vu comme un arc connectant deux nœuds (deux lexies) dans le graphe lexical du RLF ;
5. le taux de connectivité du RLF, c'est-à-dire le nombre moyen de liens de fonctions lexicales issus de chaque lexie.

En janvier 2013, soit18 mois après le début du projet RELIEF, les statistiques produites sur le RLF étaient les suivantes :

- Nombre de vocables (= V) : 11 172

- Nombre de lexies (= L) : 15 134

- Taux de polysémie (= L/V) : 1.355

- Nombre de liens de FL (= LFL) : 20 046

- Taux de connectivité (= LFL/L) : 1.325

Figure 12 : Extrait de l'entrée du vocable ADMIRER générée par l'éditeur lexicographique du RLF.

[23] ttp://www.druide.com

ADMIRER

Caractéristiques grammaticales [CG]

verbe; *de*$_{pass}$

Définition

L'individu X ~ Y pour Z(Y)
=
L'individu X apprécie Y pour Z
- beaucoup
- du fait des qualités exceptionnelles de Z

Régime

X = I = N
Y = II = N
Z = III = *pour* N, *pour* $V_{inf\text{-}passé}$, *de* $V_{inf\text{-}passé}$

Liens lexicaux

QSyn soutenu *idolâtrer*; *s'enthousiasmer*; *aimer*, *apprécier*, *estimer*

QAnti *mépriser*; *détester*; *critiquer*

Conversif

QConv$_{21}$ *plaire* [*à* N_X]

S$_0$ *admiration*

Intensément

Magn *beaucoup* < *énormément*, *profondément*; ⌜*sans réserve*⌝

Intensément et de façon exagérée

AntiBon.Magn *aveuglément*

Sans le manifester

Adv$_0$NonManif *secrètement*

Exemples

- C'est un écrivain que j'admire aussi pour son engagement politique.
- Nous ne pouvons qu'admirer l'équipe de bénévoles qui assure le bon déroulement du festival.
- Il vit en Italie, pays qu'il aime et admire.

Locutions/Clichés

- Un sot trouve toujours un sot plus sot qui l'admire

Conclusion

Le partage et la mutualisation sous le Web de versions informatiques de ressources institutionnelles sur le lexique français ouvrent des perspectives intéressantes. Ainsi le TLF, une référence en lexicographie française, a eu pendant longtemps la réputation tenace d'être un dictionnaire réservé à une élite et la diffusion de sa version papier s'est limitée à quelques milliers d'exemplaires au sein d'une intelligentsia somme toute limitée. Sa version informatique sous forme de Cédérom et d'une ressource librement accessible sur le Web a rencontré un succès important tant auprès du grand public que des utilisateurs universitaires ou des professionnels de la langue. Cette version Web fait l'objet de plus de 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 aujourd'hui un dictionnaire incontournable et un outil de promotion de la langue française.

L'intégration plus récente au sein du portail lexical du CNRTL de diverses ressources sur le lexique français permet une meilleure mutualisation des résultats de la recherche. Au-delà du seul monde universitaire, ces techniques permettent de mettre à disposition de l'ensemble de la société nos résultats de recherche. Aujourd'hui le portail lexical du CNRTL fait l'objet lui aussi de plus 400 000 requêtes par jour provenant d'horizons très divers, il est aujourd'hui l'un des sites Web sur le lexique français les plus utilisés.

Si 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, permettant un véritable travail en réseau nous avons montré qu'il convenait aujourd'hui de développer des modélisations lexicales directement intégrables dans des applications de traitement automatique des langues. Le projet RELIEF développé à Nancy s'est fixé pour objectif de développer un tel réseau lexical à large couverture afin de permettre aux ressources lexicales institutionnelles sur le français de contribuer notablement à l'amélioration des procédures de traitement automatique de notre langue.

Bibliographie

Académie française (2005), *Dictionnaire de l'Académie Française*, 9e édition, Imprimerie Nationale/ Fayard editions.

ATILF (2004) (ouvrage collectif publié sous le nom du laboratoire). *Trésor de la langue française informatisée*, Livre d'accompagnement. et CD du texte intégral (591 p.), Paris : CNRS Editions.

Bernard, P., Lecomte, J., Dendien, J. & Pierrel J.-M. (2002). Computerized linguistic resources of the research laboratory ATILF for lexical and textual analysis: Frantext, TLFi, and the software Stella, LREC 2002, *Proceedings of the Third International Conference on Language Resources and Evaluation*, (vol 3 pp. 1090-1098), Las Palmas, Espagne,.

Charest, S., Brunelle, É., Fontaine, J. & Pelletier, B. (2007). Élaboration automatique d'un dictionnaire de cooccurrences grand public. *Actes de TALN-RECITAL 2007* (vol. 1, pp 283–292), IRIT Press, Université de Toulouse-le-Mirail.

De Loupy, C., Dutoit D., El-Bèze, M. & Griot L. (1999). La partie française du lexique sémantique EuroWordNet. *Francil – Lettre d'information*, nº 17(4–5).

Dendien, J. & Pierrel, J.-M. (2003) Le Trésor de la Langue Française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence, *Traitement Automatique des Langues,* vol 44(2), 11-37.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*, Cambridge, Mass., MIT Press.

Gader, N., Lux-Pogodalla, V. & Polguère, A.(2012). Hand-Crafting a Lexical Network With a Knowledge-Based Graph Editor à paraître dans *les actes de Cogalex- 2010 : Cognitive Aspects of the Lexicon*, Post-conference workshop at COLING 2012, Mumbai, India.

Gaume, B (2004). Balades Aléatoires dans les Petits Mondes Lexicaux. In *I3: Information Interaction Intelligence*, vol. 4(2), CEPADUES édition.

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

Jousse, A.-L. & Polguère, A. (2005). *Le DiCo et sa version DiCouèbe. Document descriptif et manuel d'utilisation*. Rapport Technique, Département de Linguistique et de traduction, Université de Montréal.

Laporte, E. (1997). « Mots et niveau lexical », *Traitement Autolmatique des Langues,* vol. 38 (2).

Lux-Pogodalla, V. & Polguère, A. (2011) Construction of a French Lexical Network: Methodological Issues, *Proceedings of the First International Workshop on Lexical Resources, WoLeR 2011. An ESSLLI Workshop* (pages 54-61), Ljubljana, Slovenia,.

Martin, R., Gerner, H. & Souvay, G. (2007). DMF : Dictionnaire du Moyen Français version 2. *CILPR 2007 Congrès International de Linguistique et de Philologie Romane*, Innsbruck, Autriche.

Mel'čuk, I., Clas, A. & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*, Paris/Louvain-la-Neuve : Duculot.

Mel'čuk, I. & Polguère, A. (2007). *Lexique actif du français. L'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français (*528 p.). Louvain-la-Neuve : De Boeck,.

Mertens P. (2010). Restrictions de sélection et réalisations syntagmatiques dans DICOVALENCE. Conversion vers un format utilisable en TAL. *Actes TALN 2010,* Montréal.

Pierrel (2000). *Ingénierie des langues* (354 p.). Paris-Londres : Hermès Editions.

Pierrel, J.-M. (2008). De la nécessité et de l'intérêt d'une mutualisation informatique de connaissances sur le lexique de notre langue. 1er *Congrès Mondial de Linguistique Française*, juillet 2008, article disponible sous http://dx.doi.org/10.1051/cmlf08330.

Pierrel, J.-M. & Petitjean, É. (2007). Valorisation et exploitation scientifiques de documents numériques pour la recherche en linguistique : l'exemple du CNRTL. *Actes de CIDE 2007 Congrès International sur le Document Numérique* (pp13-24), Paris : Europia .

Poirier, C. (2005). La dynamique du français dans l'espace francophone à la lumiière de la Base de Données Lexicographiques Panfrancophone, *Revue de Linguistique Romane*, tome 69, 483-516.

Pruvost, J. (2002). *Les dictionnaires de la langue française*, collection Que Sais-je ?, Paris : PUF.

Sagot, B. & Fišer, D. (2008). Construction d'un WordNet libre du français à partir de ressources multilingues. In *TALN 2008*, Avignon.

Slodzian, M. (2000). WordNet : what about its linguistic relevancy ?, EKAW'2000, *12th international conference on knowledge engineering and knowledge management*, Juan-les-Pins.

Résumé

Au cours de la seconde moitié du 20e siècle, de nombreuses contributions institutionnelles à la lexicographie française se sont développées. Parmi elles ont peut entre autres noter : le *Trésor de la langue Française* développé par le CNRS, la 9e édition du *dictionnaire de l'Académie Française, le portail lexical du CNRTL* (Centre national de ressources textuelles et lexicales) et la *base de données lexicale panfracophone*. Ces études en lexicographie française ont donné lieu à une valorisation informatique qui a provoqué, sur le plan des études lexicales, une véritable révolution faisant de l'informatique un outil indispensable pour étudier le lexique et ses propriétés ; 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. Librement accessibles sur le Web, ces ressources ont rencontré un succès important tant auprès du grand public que des utilisateurs universitaires ou des professionnels de la langue. Référencé par d'innombrables sources, elles font l'objet de plusieurs centaines de milliers de connexions quotidiennes en provenance de tous les continents, devenant ainsi une sorte de méta-dictionnaire incontournable et un outil de promotion appréciable de la langue française.

Summary - Development and use of institutional lexical resources for French

The second half of the 20th century saw the development of many institutional contributions to French lexicography, notably the *Trésor de la Langue Française (TLF)* developed by the CNRS, the 9th edition of the *Dictionnaire de l'Académie Française, the lexical web portal* of CNRTL (National Resource centre for texts and lexica) and the *Base de Données Lexicales Panfrancophone (BDLP)*. These projects relied heavily on data-processing, thereby promoting a true revolution in lexical studies as digital tools became essential in structuring and normalizing lexicology and lexicography, leading in the process to increased mutualisation and sharing of research results for the French language, too often still widely dispersed. Freely available on the Web, these resources have met with great success both with the general public and with academics and other language professionals. They are cited in countless sources and receive several hundred thousand connections daily from all continents, becoming a kind of indispensable meta-dictionary and a tool for promotion of the French language.

Jean-Marie Pierrel
Université de Lorraine CNRS, ATILF
44, avenue de la Libération
BP 30687
54063 Nancy cedex
France
Jean-Marie.Pierrel@atilf.fr