Une nouvelle ressource lexicographique en ligne : le Petit Larousse Illustré de 1905

Helene Manuelian

Une nouvelle ressource lexicographique en ligne : le Petit Larousse Illustré de 1905

Hélène Manuélian
UMR 7187 – Lexiques, Dictionnaires et Informatique – CNRS – UCP – UP13
33 boulevard du Port
F - 95 011 CERGY PONTOISE

1. Introduction

Cet article présente une nouvelle ressource lexicographique ancienne mise à disposition sur Internet : le *Petit Larousse Illustré* de 1905. Faisant suite à des œuvres de plus grande ampleur et de plus grande renommée (le dictionnaire critique de Féraud, le dictionnaire de Nicot, les différentes éditions de celui de l'Académie, etc.), le *Petit Larousse Illustré* de 1905, bien plus modeste que ses prédécesseurs - en volume tout au moins - a été numérisé et sera mis en ligne prochainement.[1] Nous présenterons tout d'abord la ressource, les différentes étapes de sa numérisation, et enfin les différentes possibilités offertes par l'interface d'interrogation.

2. Le Petit Larousse Illustré de 1905

Le *Petit Larousse Illustré* naît en 1905, trente ans après la disparition de Pierre Larousse. Sa première édition, dirigée par Claude Augé, est mue par les mêmes idéaux de démocratisation des savoirs que ceux qui ont permis de créer les autres dictionnaires Larousse. Deux éléments font partie des raisons du succès du *Petit Larousse* : son format et ses illustrations,qui sont de deux types : les images, qui font partie de son identité et en font un objet de fascination pour les enfants en particulier ; pourtant, un autre type d'illustration participe de l'originalité et de la renommée du *Petit Larousse* : l'illustration linguistique. Le dictionnaire regorge d'exemples, puisque selon Pierre Larousse lui-même, « un dictionnaire sans exemples est un squelette ». La plupart du temps, il s'agit d'exemples forgés, qui éclairent à la fois sur le sens et l'emploi du mot, mais parfois aussi sur les aspects encyclopédiques de l'objet auquel il réfère (Pruvost, 2004). Le *Petit Larousse* est toujours, cent ans plus tard, un immense succès éditorial. C'est la première édition de cette œuvre lexicographique particulière que nous nous sommes proposé de numériser, afin de la préserver et de lui faire bénéficier des atouts de la consultation électronique.

3. Objectifs de la numérisation

3.1. □□ Préservation de l'œuvre et diffusion élargie

Le projet de numérisation du *Petit Larousse Illustré* de 1905 est né avant 2005, et donc avant la parution de son fac-similé, décidée par la maison Larousse en 2005 pour célébrer le centenaire du dictionnaire. Cela étant, malgré cette réédition, le dictionnaire du début du XX^{ème} siècle reste un objet rare, cantonné à l'espace francophone où il a été vendu, et surtout, il reste un objet de papier de plusieurs milliers de pages, ne tenant pas dans la poche... Aussi, le projet de numérisation se justifiait-il toujours pour des raisons évidentes de pérennisation du texte et de très large diffusion.

3.2. □□ Une consultation experte, permettant la recherche métalexicographique

Si la préservation de l'œuvre est un souci réel, le projet de numérisation avait surtout pour but, comme c'est souvent le cas lors de l'informatisation de telles ressources lexicales (en particulier l'informatisation de très grands dictionnaires comme celui de Féraud ou le *TLF*), la consultation experte du dictionnaire, et l'exploitation analogique que permet le format informatique (Pruvost (2000) ; Caron et al., (1996) ; Dendien et Pierrel, (2003))

[1] La mise en ligne aura lieu en juin 2010.

En premier lieu, nous souhaitions permettre une exploitation analogique du dictionnaire, grâce aux liens hypertextes. Sans aller jusqu'à utiliser des possibilités aussi importantes que l'hypernavigation dans le *TLFI*, nous permettons la navigation entre les vedettes grâce aux divers renvois (simples, synonymiques ou antonymiques) présents dans l'entrée.

De plus, nous permettons une consultation du dictionnaire, selon les champs dans l'article (certains champs ont été délimités sur la base du texte de l'article : ainsi, s'il ne contient pas explicitement une marque d'usage, nous n'en ajoutons pas, même s'il s'agit d'un oubli).

4. Utilisation de la TEI et exploitation d'une analyse lexicographique minutieuse

4.1. Utilisation des standards informatiques liés aux ressources textuelles

Pour être diffusé et consulté de façon optimale, et être conservé le plus longtemps possible, la base doit répondre à des standards informatiques. Nous avons donc opté pour une numérisation dans un format XML, et un balisage conforme à la TEI P5 (Ide et Véronis (1994) ; TEI Consortium). Cela étant, le dictionnaire est ancien, pas toujours homogène, et certains éléments qui apparaissent dans les entrées ont parfois nécessité le typage des balises existantes. Ainsi par exemple, les indications d'étymologie contiennent parfois des données morphologiques. C'est la raison pour laquelle nous avons introduit un type « morpho » pour distinguer les notes purement étymologiques, des indications morphologiques, comme le montre l'exemple suivant :

```
ABLÉGAT (…) <etym type=morpho>(préf. <mentioned>ab</mentioned>, et
<lang>lat.</lang> <mentioned>legatus</mentioned>, <gloss>envoyé</gloss>)</etym>
(…).
ABOI (…) <etym type=morpho> (de <i>aboyer</i>) </etym> (…).
```

Malgré ces difficultés, le balisage a été rendu possible grâce à deux choses : une analyse lexicographique minutieuse et une utilisation précise[2] de la forme du texte par des programmes informatiques repérant ainsi l'organisation des entrées.

4.2. Une analyse lexicographique minutieuse

Afin de pouvoir écrire une DTD et réaliser le balisage, le dictionnaire a été étudié minutieusement par Jean Pruvost, qui a analysé chaque article, et typé chaque exemple du dictionnaire. L'analyse était bien entendu si minutieuse qu'il n'a pas été toujours possible de la restituer dans sa précision au balisage, mais c'est bien elle qui a permis le repérage des divers éléments du dictionnaire.

4.3. Utilisation de la forme pour poser des balises sémantiques

Grâce à l'analyse des articles, nous avons constaté que la forme typographique des entrées est d'une telle régularité, qu'on devrait pouvoir, grâce aux indications de police et à la position relative des différents composants de l'entrée, pouvoir écrire des programmes en langage *Python* utilisant des expressions régulières qui baliseront récursivement les articles du dictionnaire. Pour plus de clarté, nous développons notre idée sur un exemple.

Voici l'article du *Petit Larousse Illustré* de 1905 pour le mot *aile* tel qu'il apparaît dans le dictionnaire[3].

> **AILE** *(è-le)* n. f. (lat. *ala).* Membre des oiseaux et de quelques insectes, qui leur sert à voler. *Par ext. Ailes d'un moulin,* ses châssis garnis de toiles. *Ailes d'un bâtiment,* ses côtés. *Ailes d'une armée,* ses flancs. *Fig.* Protection, surveillance : *se réfugier sous*

[2] Bien sûr, le dictionnaire est une œuvre humaine, et tout n'y est pas régulier ; de nombreuses erreurs de balisage liées à des erreurs ou irrégularités dans le document d'origine ont été constatées, contournées ou ont nécessité de nombreuses corrections manuelles du balisage.

[3] Cette entrée est choisie de façon à montrer la richesse des informations contenues dans le dictionnaire, leur nombre et leur enchaînement. La plupart des entrées sont bien entendu généralement plus courtes.

l'aile de sa mère. LOC. PROV**. : Voler de ses propres ailes,** se passer d'autrui. **Battre de l'aile**, être embarrassé, mal à l'aise. **Rogner les ailes à quelqu'un,** lui retrancher de son autorité, de son revenu. Tirer une plume de l'aile à quelqu'un, lui attraper quelque chose, lui extorquer de l'argent. **A tire-d'aile.** V. TIRE D'AILE.

La mise en forme est valable pour toutes les entrées et nous la décrivons ainsi :
- la vedette est en majuscule et en gras ; la prononciation apparaît derrière la vedette en italiques (mais n'est pas présente dans toutes les entrées) ;
- les informations grammaticales apparaissent abrégées derrières la note de prononciation les marques d'usage (domaine, style, registre…) sont données en abrégé et en italiques.
- les informations étymologiques sont entre parenthèses derrière l'information grammaticale et certains éléments contenus dans les parenthèses (les noms des langues) constituent eux aussi une liste fermée ;
- les définitions apparaissent en caractères droits derrière l'étymologie, une marque d'usage, une expression, une locution proverbiale, un proverbe ou une forme dérivée et finissent par un point ;
- les exemples apparaissent après deux points (:) et sont en italiques ;
- les expressions et la phraséologie de façon générale apparaissent après une marque d'usage ou une définition, en italiques et finissent par une virgule ;
- les renvois (simples, synonymiques ou antonymiques) sont en petites majuscules ;
- les proverbes ou locutions sont signalés explicitement en petites majuscules, apparaissent en gras et finissent par une virgule ;

Le dictionnaire est numérisé, puis converti au format HTML, ce qui permet de travailler sur un format standard et surtout, sur un balisage formel des entrées.

```
<p><b>AILE</b> (<i>è-le</i>) n. f. (lat. <i>ala</i>). Membre des oiseaux et <lb>de
quelques insectes, qui leur sert à voler. <i>Par ext. <lb>Ailes d&#39;un
moulin</i>, ses châssis garnis de toiles. <i>Ailes <lb>d&#39;un bâtiment</i>, ses
côtés. <i>Ailes d&#39;une armée</i>, ses <lb>flancs. <i>Fig</i>. Protection,
surveillance : <i>se réfugier sous <lb>l&#39;aile de sa mère</i>. <span
style="font-variant: small-caps">Loc. prov</span>. <b>: Voler de ses propres
<lb>ailes</b>, se passer d&#39;autrui. <b>Battre de l&#39;aile</b>, être
em<lb>barrassé, mal à l&#39;aise. <b>Rogner les ailes à quel<lb>qu&#39;un</b>,  lui
retrancher de son autorité, de son revenu. <lb> <b>Tirer une plume de l&#39;aile à
quelqu&#39;un</b>,  lui attrap<lb>per quelque chose, lui extorquer de l&#39;argent.
<b>A tire-<lb>d&#39;aile</b>.  V. <span style="font-variant: small-caps">tire-
d&#39;aile</span>.</p>
```

Les fichiers sont ensuite convertis en XML, puis les programmes sont passés de façon à baliser automatiquement les différents éléments. Comme ils utilisent les positions relatives des différents éléments, leur ordre est crucial, et le balisage se déroule en trois passes.
La première repère les entrées, les informations grammaticales, les prononciations et les marques d'usage. L'entrée balisée prend alors la forme suivante :

```
<EntryFree><form><orth>AILE</orth> <pron>(è-le)</pron></form>
<gramGrp><pos>n.</pos> <gen>f.</gen></gramGrp> (lat. <i>ala</i>). Membre des
oiseaux et de quelques insectes, qui leur sert à voler. <i><usg type="style">Par
ext.</usg> Ailes d&#39;un moulin</i>, ses châssis garnis de toiles. <i>Ailes
d&#39;un bâtiment</i>, ses côtés. <i>Ailes d'une armée</i>, ses flancs. <usg
type="style">Fig.</usg> Protection, surveillance : <i>se réfugier sous l'aile de sa
mère.</i> <span style="font-variant: small-caps">Loc. prov</span>. <b>: Voler de
ses propres ailes</b>, se passer d'autrui. <b>Battre de l'aile</b>, être
embarrassé, mal à l'aise. <b>Rogner les ailes à quelqu'un</b>,  lui retrancher de
son autorité, de son revenu. <b>Tirer une plume de l'aile à quelqu'un</b>,  lui
attraper quelque chose, lui extorquer de l'argent. <b>A tire-d'aile</b>. V. <span
style="font-variant: small-caps">tire-d'aile</span>.</EntryFree>
```

La deuxième passe repère les notes étymologiques, les renvois, les proverbes et les exemples :

```
<EntryFree><form><orth>AILE</orth> <pron>(è-le)</pron></form>
<gramGrp><pos>n.</pos> <gen>f.</gen></gramGrp> <etym>(<lang>lat.</lang>
<mentioned>ala</mentioned>)</etym>. Membre des oiseaux et de quelques insectes, qui
leur sert à voler. <usg type="style">Par ext.</usg> <i>Ailes d'un moulin</i>, ses
châssis garnis de toiles. <i>Ailes d'un bâtiment</i>, ses côtés. <i>Ailes d'une
armée</i>, ses flancs. <usg type="style">Fig.</usg> Protection, surveillance <cit
type="example">: <quote>se réfugier sous l'aile de sa mère.</quote></cit> L<sc>oc.
prov</sc>. <b>: Voler de ses propres ailes</b>, se passer d'autrui. <b>Battre de
l'aile</b>, être embarrassé, mal à l'aise. <b>Rogner les ailes à quelqu'un</b>, lui
retrancher de son autorité, de son revenu. <b>Tirer une plume de l'aile à
quelqu'un</b>, lui attrapper quelque chose, lui extorquer de l'argent. <b>A tire-
d'aile</b>. <xr type="renvoi"><lbl>V.</lbl> <ref target ="#tire-d'aile">tire-
d'aile</ref></xr>.</EntryFree>
```

La troisième passe repère les définitions, et les sous – entrées (dérivés, phraséologie) :

```
<EntryFree><form><orth>AILE</orth> <pron>(è-le)</pron></form>
<gramGrp><pos>n.</pos> <gen>f.</gen></gramGrp> <etym>(<lang>lat.</lang>
<mentioned>ala</mentioned>)</etym>. <def>Membre des oiseaux et de quelques
insectes, qui leur sert à voler</def>. <usg type="style">Par ext.</usg> <re
type="exp"><form>Ailes d'un moulin</form>, <def>ses châssis garnis de
toiles</def></re>. <re type="exp"><form>Ailes d'un bâtiment</form>, <def>ses
côtés</def></re>. <re type="exp"><form>Ailes d'une armée</form>, <def>ses
flancs</def></re>. <usg type="style">Fig.</usg> <def>Protection, surveillance</def>
<cit type="example">: <quote>se réfugier sous l'aile de sa mère.</quote></cit> <cit
type="Loc. Prov.">Loc. Prov.<quote>  Voler de ses propres ailes</quote></cit>,
<def>se passer d'autrui</def>. <cit type="Loc. Prov."><quote>Battre de
l'aile</quote></cit>, <def>être embarrassé, mal à l'aise</def>. <cit type="Loc.
Prov."><quote>Rogner les ailes à quelqu'un</quote></cit>, <def>lui retrancher de
son autorité, de son revenu</def>.. <cit type="Loc. Prov."><quote>Tirer une plume
de l'aile à quelqu'un</quote></cit>, <def>lui attrapper quelque chose, lui
extorquer de l'argent</def>. <def>A tire-d'aile</def>. <xr
type="renvoi"><lbl>V.</lbl> <ref target ="#tire-d'aile">tire-
d'aile</ref></xr>.</EntryFree>
```

Nous ne détaillerons pas dans cet article les phases intermédiaires de relectures et autres conversions qui ont ponctué le processus. Ce qui nous importe ici, c'est de montrer que l'exploitation de la forme du texte et un recensement minutieux des éléments qui le composent ont permis son balisage sémantique, respectant (à quelques exceptions près) les recommandations de la TEI.

5. Le résultat : une interface d'interrogation experte de la ressource

5.1. Une recherche plein texte classique

L'interface d'interrogation va permettre une recherche plein texte des plus classique. L'utilisateur saisit un mot de son choix dans la fenêtre d'interrogation, et l'interface affiche la totalité de l'entrée contenant le mot recherché.

5.2. Une interrogation avancée grâce à l'analyse du dictionnaire

Nous avons par ailleurs souhaité une interrogation experte de la base de données. En effet, l'analyse du dictionnaire étant fine, et le balisage ayant essayé de rendre l'essentiel de cette analyse, nous pensions qu'il était plus que nécessaire de permettre à l'utilisateur d'en bénéficier. Nous allons donc pouvoir interroger la plupart des champs balisés, de façon différente en fonction du type de champ étudié.

Les vedettes, définitions, proverbes, et exemples fonctionnent de la même manière : l'utilisateur saisit un mot librement, et les entrées correspondant à la demande s'affichent.

D'autres champs, correspondant à des listes fermées de valeurs, ne permettent pas une saisie libre, mais proposent une liste de valeurs possibles. Ainsi, par exemple, si l'utilisateur fait une recherche sur la langue d'origine des mots mentionnée dans la note étymologique, il choisit dans le menu déroulant correspondant aux champs du dictionnaire « note étymologique – langue » puis entre les différentes langues proposées par le dictionnaire. Là encore nous ne nous attarderons pas sur les détails techniques, mais tenons à signaler une facilité offerte à l'utilisateur. Certaines langues apparaissent abrégées de façons très variées en fonction des contextes. Ainsi, une langue comme l'ancien haut allemand peut être mentionnée comme

« anc. h. allem. » ou « anc. haut allem. » ou encore « anc. h. all. » etc. Elle n'apparaîtra que sous la forme entière dans l'interface[4].

5.3. Exemples de résultats d'interrogation de la base

La base de données présente en résultats des requêtes les entrées complètes. Le résultat de la recherche plein texte et celui de la recherche sur les vedettes offre la possibilité de voir les images associées aux entrées dans le dictionnaire d'origine. Les copies d'écran présentent respectivement l'interface de recherche, avec la possibilité de recherche plein texte et la possibilité de recherche avancée (Figure 1), le résultat de la recherche plein texte sur le mot abeille (Figure 2) et le résultat d'une recherche sur les mots provenant de l'ancien haut allemand (Figure 3)[5].

Figure 1 : Interface d'interrogation de la base

Figure 2 : Résultat de la recherche plein texte du mot « abeille »

[4] Ceci est valable pour tous les champs dont les valeurs ont été recensées dans une liste fermée d'items : catégorie grammaticale, langue d'origine, marques d'usage.

[5] Les résultats présentés ne sont que partiels, puisqu'à la date où nous rédigeons cet article, seules les lettres A, B et C sont totalement achevées (soit 25% du dictionnaire).

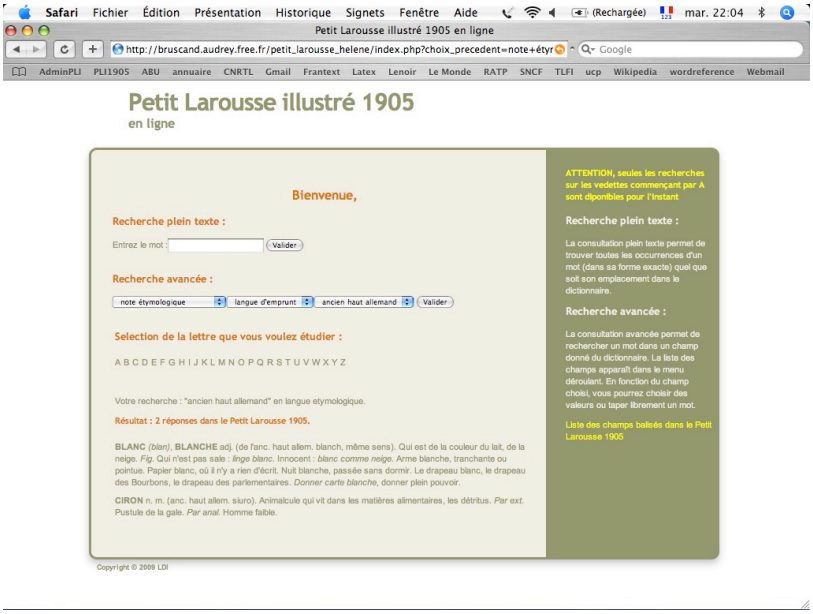

Figure 3 : Résultat de la recherche sur les mots ayant l'ancien haut allemand pour origine

6. Conclusion

Le *Petit Larousse* de 1905 informatisé est donc une nouvelle ressource lexicale en ligne, disponible librement et sans abonnement, pour la communauté scientifique comme pour l'utilisateur lambda. Elle présente comme intérêt de reproduire fidèlement le dictionnaire de 1905 (à terme, avec les illustrations) et de permettre une interrogation experte de la base de données.

Elle permet aussi de démontrer l'adaptation (moyennant des aménagements, mais bien réelle) de la TEI P5 à une ressource lexicographique pensée bien avant l'invention de l'informatique, ce qui nous semble important, étant donné que parfois, les lexicographes vivent les standards informatiques comme des contraintes, et non comme des aides ou des observations.

7. Bibliographie

Caron P., Dagenais L., Gonfroy G. (1996). « Le programme d'informatisation du "Dictionaire critique de la langue française" de l'abbé Jean-François Féraud (1787) », CHWP B.6. Mai 1996.

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. 44 - 2

Pruvost J. (2000). *Dictionnaires et nouvelles technologies*, Paris : Presses universitaires de France.

Pruvost J. (2004). *La dent-de-lion, la semeuse et le Petit Larousse*, Paris : Larousse.

TEI consortium, *Print Dictionaries*, TEI P5, http://www.tei-c.org/release/doc/tei-p5-doc/html/DI.html