Presenting the Nénufar Project: a Diachronic Digital Edition of the Petit Larousse Illustré

Hervé Bohbot, Francesca Frontini, Giancarlo Luxardo, Mohamed Khemakhem, Laurent Romary

Presenting the *Nénufar Project*: a Diachronic Digital Edition of the *Petit Larousse Illustré*

Hervé Bohbot, Francesca Frontini, Giancarlo Luxardo
PRAXILING UMR 5267 Univ Paul Valéry Montpellier 3 & CNRS - Montpellier, France
name.surname@univ.montp3.fr

Mohamed Khemakhem[1,2,3], Laurent Romary[1,2,4]
[1] Inria – ALMAnaCH, Paris
[2] Centre Marc Bloch, Berlin
[3] Université Paris Diderot, Paris
[4] Berlin-Brandenburgische Akademie der Wissenschaften, Berlin
name.surname@inria.fr

Abstract

This paper presents the Nénufar project, which aims to make several successive (free of copyright up to 1948) editions of the **French Petit Larousse Illustré** dictionary available in a digitised format. The corpus of digital editions will be made publicly available via a web-based querying interface, as well as distributed in a machine readable format, TEI-LEX0.

Keywords: TEI, Petit Larousse, dictionaries

1. Introduction

The digitisation of historical dictionaries has recently taken on strong momentum, moving past the mere publication of scanned texts to the conversion of paper dictionaries into easily exploitable lexical databases encoded using well established digital standards. At the same time, a number of the main historical French dictionaries (16th to 19th century) are also currently being digitised and made available online. Two main initiatives in this regard are *Grand Corpus des dictionnaires Garnier*[1] and the ARTFL project[2], which provide access to the content by means of search interfaces (though access is partly restricted and sources aren't downloadable)[3]. On the other hand there is a lack of similar initiatives for 20th century French dictionaries.

The **Nénufar**[4] **project** aims to make several successive editions of the *Petit Larousse Illustré* (PLI) available in a digitised format. The PLI makes an especially good candidate for such a project since it is the only French dictionary that has been updated every year since it was first published, in this case in 1905. Under the French copyright law, collective works such as the PLI fall under the public domain after 70 years from the publication, which means that we can at present take into account all editions up to 1948.

Each new edition of the PLI differs from the previous one in terms of lexical entries (with a number of words entering or exiting); but changes are also found in updated definitions and at times in the orthographic and grammatical norms which are referred to, all of which provides lexicographers, linguists and historians with an invaluable source of information on the evolution of French language and culture during the first half 20th century . At the same time, the evolution of language notwithstanding, the PLI is also an important source of linguistic information on contemporary French, and its digitisation will feed into the existing ecosystem of French language technologies (see (Mariani et al., 2012) for an overview).

2. The Project

Nénufar is a project headed by headed by laboratoire Praxiling at the Paul Valéry University of Montpellier in collaboration with INRIA, and is supported by funding from the Delegation Generale a la Langue Francaise et aux Langues de France (DGLFLF) and the Huma-Num consortia CORLI[5] and CAHIER[6]. It continues a previous project initiated in the early 2000s and which saw the publication a first version of the 1905 edition in 2005[7].

The original edition was available for searching from a web interface, which is no longer available; moreover, the XML encoding used is not fully TEI compliant.

The first goal of the Nénufar project is thus to re-encode the 1905 edition, transforming the existing version into a TEI compliant XML, as well as correcting remaining OCR errors and improving the detection and annotation of the main lexicographic elements of each entry.

The availability of an already existing digitised version of the first edition makes the digitisation of later editions much easier: by comparing two OCRed versions of two subsequent editions it is possible to identify changes in the more recent edition, but also undetected OCR errors from the previous one.

While the PLI was published every year since 1905 the project will prioritise the digitisation of only a selected set

[1] http://www.classiques-garnier.com/

[2] http://artfl-project.uchicago.edu

[3] Gallica also provides access to OCRed scans of old dictionaries, http://gallica.bnf.fr/.

[4] (Nouvelle édition numérique de fac-similés de référence)

[5] https://corli.huma-num.fr/

[6] http://cahier.hypotheses.org/

[7] This first initiative was headed by laboratoire Lexique, Dictionnaires et Informatique, under the lead of Jean Pruvost, who is now an advisor in Nénufar.

of issues, which correspond to major re-editions of the dictionary - namely the 1924, 1936, 1948 ones.
Currently the 1924 edition is being digitised, and we calculated that 1/3 of its entries were modified with respect to the 1905 one.
A first release of the Nénufar corpus, including the 1905 and the 1924 editions, will take place by the end of 2018. New editions will be subsequently made available. Alongside with the lexicographic part, it will also contain additional onomastic information (from the encyclopedic section of the PLI, listing proper names of people, places, ...) and a digitised version of all figures with their captions.

3. The Formats

The question of publication formats is crucial for a project such as this one, which caters to different research communities. On the one hand, in order to fit the requirements of the general public as well as of traditional historical lexicographers, we need to provide a browsable web interface, which enables users to search for entries and see their evolution over time in a user-friendly way. On the other hand, the needs of digital lexicographers and language technologists can only really be met by making the sources of each edition available in a standardised format, something that would not only allow for more specialised querying, but would also be best suited for long term preservation.
Currently two formats are under discussion for the publication of retrodigitised dictionaries such as PLI, namely the TEI dictionaries module[8], the Ontolex-Lemon model (RDF) (McCrae et al., 2017). Those two formats serve different purposes: TEI represents the dictionary as a digital edition, and is better suited to the needs of lexicographers and linguists, while Ontolex-Lemon is the reference format for the publication of dictionaries as Linked Open Data, and thus is more relevant for the domain of Language and Semantic Web technologists.
As to the encoding of PLI in TEI, the first step was to transform the 2005 mark-up in a TEI compliant format, which is the one presented in Appendix B. This first encoding remains very adherent to the structure of the typographic entry, as can be seen in Appendix A, and thus uses the *entryFree* TEI tag, which allows for maximum freedom in the representation and encoding of the different parts of a lexical entry. For this reason it is the one that will be used internally in the Nénufar database to derive the HTML displayed on the browsable web interface.
However an excessive freedom in terms of entry modelling can become a hindrance to interoperability with other projects. For this reason a recent a joint ENeL[9] / DARIAH[10] / PARTHENOS[11] initiative has proposed a more strict TEI representation for dictionaries, called TEI-Lex0(Bański et al., 2017). TEI-Lex0 derives from the lexicographic module of TEI and is fully TEI compliant, but aims to provide more clear guidelines for the encoding of retrodigitised dictionaries.
With respect to the more general TEI guidelines for dictionaries, TEI-Lex0 is aimed at providing a schema which will allow most modern dictionaries to be represented in a way that enables interoperability, comparability and further ease of exploitation. To that end, the internal structure and information of lexical entries have been revised and optimised to be more clearly explicit and uniform.
We believe that the PLI can constitute an excellent test case for this new format, which we intend as the distribution format for the downloadable resource. In Appendix C you can find the same entry transformed into the TEI-Lex0 format.
As you can see, going from the current format to the new one requires some changes; some of them (such as the insertion of the *type* attribute in the *form* tag) are straightforward, but others are more complex to implement.
Firs of all the *entryFree* tag is replaced by *entry*, which allows for less freedom as to the tags it may contain. As a consequence, the original structure cannot be left as it is. In particular the *sense* tag needs to be inserted, to group a definition with its related examples and citations. This implies adding information which, in the original entry is not explicitly marked by visible typographic features (such as numbering, symbols or formatting, as is the case in other dictionaries). By close analysis of the PLI entries, we consider that every new definition instantiates a new sense, and that no sense hierarchy is inferable.
Another issue is the fact that free text is not allowed within the *sense* tag. Thus *pc* tags need to be used to wrap up punctuation elements such as columns, as they cannot be considered neither as part of the definition, nor of the citation.
Despite the work required to transform the current format into TEI-Lex0, the advantages are obvious; TEI-Lex0 will allow for different dictionaries to be queried using the same strategy and also facilitate the development of common tools.
One of the current applications of this format is in the GROBID-Dictionaries infrastructure, which aims to automatically machine-learn the TEI-Lex0 structure of a dictionary entry from OCRed dictionary pages (Khemakhem et al., 2017). Within the Nénufar project experiments are ongoing to digitise new editions with GROBID-Dictionaries.
As to the Ontolex-Lemon version, at the time of writing this paper (March 2018) a working group is active drafting the specifications for a dictionary module, which will enable to represent retro-digitised dictionaries using the Ontolex-Lemon core with additional properties. The specifications are not yet finalised, and the final modelling of PLI in this new format will be the object of further research; it is important however to underline how PLI entries from the 1905 edition are currently being used as examples to discuss the new module issues[12].
As to the availability of the two versions, the TEI edition will be downloadable from the Ortolang[13] platform, and the Ontolex-Lemon will be queryable via a SPARQL endpoint.
Finally, two modelling issues are of a more generic nature and will affect both formats. On the one hand homographs

[8](Budin et al., 2012), see also `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html`
[9]`http://www.elexicography.eu/`
[10]`https://www.dariah.eu/`
[11]`http://www.parthenos-project.eu/`
[12]`https://www.w3.org/community/ontolex/wiki/Lexicography`
[13]`http://www.ortolang.fr`

are generally but not systematically treated as separate entries in the PLI; this may represent a problem as to the encoding of grammatical properties at the entry level and may require adjustments. On the other a normalisation of data categories for grammatical features is required and currently on-going; the grammatical labels (gender, number, language, ...), represented with in the original by (often unsystematic) French abbreviations, will be normalised using existing controlled vocabularies; in this sense, the CLARIN Concept Registry may [14] constitute a valid solution.

4. The Content

Dictionaires are the "tools of a language and a culture" (Pruvost, 2006) and the PLI, whose hundreds of thousands of copies reached the majority of French households, has played and still plays a great role in the democratisation of linguistic knowledge(Cormier et al., 2006); for this reason the diachronic investigation of its successive editions sheds a new light on the evolution of French language and society.
First and foremost the Nénufar corpus will constitute a privileged source of information on the evolution of orthography. The name of the project itself is inspired by a surprising controversy sparked in 2016 by the proposed change in the spelling of the French word for waterlily, from *nénuphar* to *nénufar*. Despite the fact that the new spelling was strongly ostracised by the people and by the media, an inspection of early editions of PLI shows that the *nénufar* spelling was already present in the 1905 edition and remained the preferred orthography for the word for the whole of the first half of the 20th century. Other orthographies attested in the earlier versions PLI would be considered almost shocking today, such as *à priori* (with an accent), *fiord* instead of *fjord*, *ognon* as an alternate spelling for *oignon* (the French for *onion*).
Apart from the evolution of orthography, the older editions of the PLI are rich in information about phonetics ([distrik], [lo-kouass] for *district* et *loquace* en 1906), neologisms (*antimilitarisme* in 1911, *boche*, the equivalent of the English pejorative word for German, in 1917, etc.) and changes in the definitions. As to the these, some are rather amusing, such as the one for *aviation*, which in 1905 reads "on a fait de nombreuses tentatives à ce sujet mais le problème n'est pas encore résolu" (several tests have been carried out but the problem hasn't been solved yet) and in 1911 becomes "les aéroplanes ont victorieusement résolu le problème du plus lourd que l'air" (planes have victoriously solved the heavier-than-air controversy). In other cases (as in the older entries for *juiverie* or *nègre, négresse*) definitions bear testimony of the evolution of society, of which the PLI is the mirror.

5. Conclusion

In this paper we presented Nénufar, an ongoing project aimed to the digitisation of chosen editions of the Petit Larousse Illustré from the first half of the 20th century.
A first TEI and web release of the Nénufar corpus will be available in 2018 with an open license, thus enabling research in the domains of linguistics, history and language technologies to research and use this
To ensure interoperability, the project is carried out in close contact with on-going international initiatives aimed at promoting standard and best practices in the retro-digitisation of legacy dictionaries[15]. Moreover, it is currently used as a test bed for GROBID-Dictionaries, a technology which will considerably speed up the encoding of OCR-ed resources.
The current project is specifically targeting the PLI, but the best practices developed within Nénufar will be applicable to other legacy dictionaries.

6. Bibliographical References

Bański, P., Bowers, J., and Erjavec, T. (2017). TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms. In *eLex2017*.

Budin, G., Majewski, S., and Mörth, K. (2012). Creating Lexical Resources in TEI P5. *Journal of the Text Encoding Initiative*, (Issue 3), November.

Cormier, M.-C., Pruvost, J., Mitterand, H., Garnier, Y., and Collectif. (2006). *Les dictionnaires Larousse : Genèse et évolution*. PU Montréal, Montréal, March.

Khemakhem, M., Foppiano, L., and Romary, L. (2017). Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. In *electronic lexicography, eLex 2017*, Leiden, Netherlands, September.

Joseph Mariani, et al., editors. (2012). *La langue française à l' Ère du numérique – The French Language in the Digital Age*. White Paper Series. Springer-Verlag, Berlin Heidelberg.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In *eLex2017*.

Pruvost, J. (2006). *Les dictionnaires français : Outils d'une langue et d'une culture*. Ophrys, Paris.

[14] https://concepts.clarin.eu/ccr/browser/

[15] In addition to what was mentioned in this paper, Nénufar is planning on collaborating with the ELEXIS project, which recently kicked off and aims at building a European Infrastructure for E-lexicography (http://www.elex.is/)

Appendices

A The dictionary entry *verre* (glass) in the PLI .

VERRE (*vè-re*) n. m. (lat. *vitrum*). Corps solide, transparent et fragile, produit de la fusion d'un sable siliceux mêlé de potasse ou de soude : *le verre est très cassant*. Objet fait de verre : *verre de montre*. Vase à boire, fait de verre ; ce qu'il contient : *un verre de vin*. *Verre double*, verre très épais. *Maison de verre*, maison où il n'y a rien de secret. *Petit verre*, liqueur alcoolique qu'on prend dans un verre de petite dimension : *boire un petit verre*. — Le *verre*, dont l'invention est attribuée aux Phéniciens, est obtenu par la fusion dans des *creusets* (ou *pots*) d'un mélange de silice (sable) avec des sels de soude, de potasse (*verre ordinaire*) ou de plomb (*cristal*). Les creusets sont placés dans des *fours* où la température est poussée jusqu'à 1.000°. Cueilli avec une *canne* que l'on plonge dans les creusets par une ouverture (*ouvreau*) pratiquée dans la paroi du four, le verre pâteux est travaillé, soufflé, moulé, étiré, pour donner des bouteilles, des vitres, des objets de gobeleterie, des tubes, etc. Les glaces sont obtenues par *coulage ;* on sort du four le creuset tout entier et l'on en verse le contenu sur une immense table de fonte. Tous les objets de verre, avant d'être livrés au commerce et indépendamment des façons qu'on leur fait subir ou des décors dont on les agrémente, doivent être *recuits* c'est-à-dire refroidis lentement, pour être moins cassants. Outre les mille objets à l'usage domestique, le verre sert encore à fabriquer les verres optiques et les instruments si nombreux utilisés dans les laboratoires. Ramolli au four et comprimé fortement, il donne la *pierre de verre*, qu'on emploie au revêtement des murs et même au pavage des rues.

Véronique.

B The first TEI-XML encoding

```
<entryFree xml:id="verre">
  <form>
    <orth>VERRE</orth>
  </form>
  <pron>(vè-re)</pron>
  <gramGrp>
    <pos>n.</pos>
    <gen>m.</gen>
  </gramGrp>
  <etym>
    (<lang>lat.</lang> <mentioned>vitrum</mentioned>)
  </etym>.
  <def>Corps solide, transparent et fragile, produit de la fusion d'un sable
    siliceux mêlé de potasse ou de soude</def> :
  <cit type="example"><quote>le verre est très cassant.</quote></cit>
  <def>Objet fait de verre</def> :
  <cit type="example"><quote>verre de montre.</quote></cit>
  <def>Vase à boire, fait de verre ; ce qu'il contient</def> :
  <cit type="example"><quote>un verre de vin.</quote></cit>
  <re type="exp"><form>Verre double</form>, <def>verre très épais.</def></re>
  <re type="exp"><form>Maison de verre</form>,
    <def>maison où il n'y a rien de secret.</def>
  </re>
  <re type="exp"><form>Petit verre</form>,
    <def>liqueur alcoolique qu'on prend dans un verre de petite dimension</def> :
    <cit type="example"><quote>boire un petit verre.</quote></cit>
  </re> -
  <def value="encycl">
    Le <emph rend="italic">verre</emph>, dont l'invention est attribuée
    aux Phéniciens, est obtenu par la fusion dans des <emph rend="italic">
      creusets</emph> (ou <emph rend="italic">pots</emph>) d'un mélange de
    silice (sable) avec des sels de soude, de potasse (<emph rend="italic">
      verre ordinaire</emph>) ou de plomb (<emph rend="italic">cristal</emph>.)
    Les creusets sont placés dans des <emph rend="italic">fours</emph> où la
    température est poussée jusqu'à 1.000°. Cueilli avec une <emph rend="italic">
      canne</emph> que l'on plonge dans les creusets  par une ouverture
    (<emph rend="italic">ouvreau</emph>) pratiquée dans la paroi du four,
    le verre pâteux est travaillé, soufflé, moulé, étiré, pour donner des
    bouteilles, des vitres, des objets de gobeleterie, des tubes, etc.
    Les glaces sont obtenues par <emph rend="italic">coulage</emph> ;
    on sort du four le creuset tout entier et l'on en verse le contenu
    sur une immense table de fonte. Tous les objets de verre, avant d'être
    livrés au commerce et indépendamment des façons qu'on leur fait subir
    ou des décors dont on les agrémente, doivent être
    <emph rend="italic">recuits</emph> c'est-à-dire refroidis lentement,
    pour être moins cassants. Outre les mille objets à l'usage domestique,
    le verre sert encore à fabriquer les verres optiques et les instruments
    si nombreux utilisés dans les laboratoires.
    Ramolli au four et comprimé fortement,
    il donne la <emph rend="italic">pierre de verre</emph>, qu'on emploie
    au revêtement des murs et même au pavage des rues.
  </def>
</entryFree>
```

C The TEI-LEX0 encoding

```
<entry>
  <form type="lemma">
    <orth>VERRE</orth>
    <pron>(vè-re)</pron>
  </form>
  <gramGrp>
    <pos>n.</pos>
    <gen>m.</gen>
  </gramGrp>
  <etym> (<lang>lat.</lang>
    <mentioned>vitrum</mentioned>)</etym><pc> .</pc>
  <sense>
    <def>Corps solide, transparent et fragile, produit
      de la fusion d'un sable siliceux
      mêlé de potasse ou de soude</def><pc> :</pc>
    <cit type="example">
      <quote>le verre est très cassant</quote>
    </cit>
  </sense><pc>.</pc>
  <sense>
    <def>Objet fait de verre</def><pc> :</pc>
    <cit type="example">
      <quote>verre de montre</quote>
    </cit>
  </sense><pc>.</pc>
  <sense>
    <def>Vase à boire, fait de verre ;
      ce qu'il contient</def><pc> :</pc>
    <cit type="example">
      <quote>un verre de vin</quote>
    </cit>
  </sense><pc>.</pc>
  <re type="exp">
    <form type="lemma">
      <orth>Verre double</orth>
    </form>, <def>verre très épais</def></re><pc>.</pc>
  <re type="exp">
    <form type="lemma">
      <orth>Maison de verre</orth>
    </form>, <def>maison où il n'y a rien de
      secret</def>
  </re><pc>.</pc>
  <re type="exp">
    <form type="lemma">
      <orth>Petit verre</orth>
    </form><pc>,</pc>
    <sense>
      <def>liqueur alcoolique qu'on prend dans un
        verre de petite dimension</def><pc> :</pc>
      <cit type="example">
        <quote>boire un petit verre</quote>
      </cit>
    </sense>
  </re><pc>.</pc>
  <sense>
    <def value="encycl"> - Le <emph rend="italic">verre</emph>, dont l'invention
      est attribuée aux Phéniciens, [...see encoding above....]</def>
  </sense>
</entry>
```