A Diachronic Digital Edition of the Petit Larousse illustré

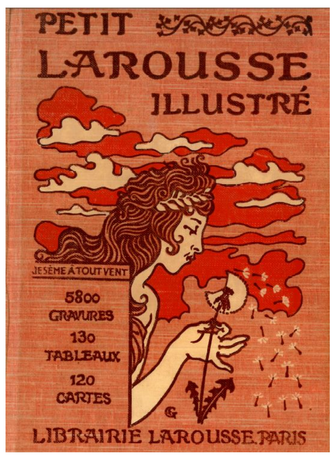
Herve Bohbot, Alexandre Faucher, Francesca Frontini, Agata Jackiewicz, Giancarlo Luxardo, Agnès Steuckardt, Mohamed Khemakhem, Laurent Romary

Presenting the Nénufar project:

A Diachronic Digital Edition of the *Petit Larousse illustré*

Hervé Bohbot, Alexandre Faucher, Francesca Frontini, Agata Jackiewicz, Giancarlo Luxardo, Agnès Steuckardt
Praxiling UMR 5267, CNRS – Univ. Paul-Valéry Montpellier 3, Montpellier, France

Mohamed Khemakhem, Laurent Romary
Inria – ALMAnaCH, Paris – Centre Marc Bloch, Berlin – Université Paris Diderot, Paris
Berlin-Brandenburgische Akademie der Wissenschaften, Berlin

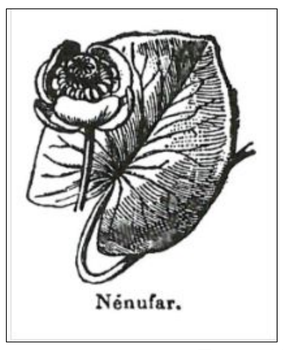


- The *Petit Larousse illustré* is the only French dictionary to be updated every year, since it was first published in 1905 (coined *1906*).
- Handy and affordable, with each edition selling several hundreds of thousands copies (Pruvost, 2002), one *Petit Larousse* may virtually be found in every French household.
- Editions published prior to 1948 are public domain according to French laws, however, no open full text resource is currently available.

Language section (1906):
- 1066 pages *in-8°*
- 44876 entries
- 2532 illustrations
- ca 1.06 million words

Nénufar project: aiming to a diachronic corpus of 1906 to 1948 editions

Nénufar project name stands for *Nouvelle édition numérique de fac-similés de référence*. This name was chosen after a typically French polemic between supporters and opponents of some orthographic modifications of the French language. Amongst others, it was proposed that the waterlily (*Nymphæa* L.), should be spelled *nénufar* instead of *nénuphar*, which was perceived by some conservative people as an aggression towards the language integrity… However, it *was actually* the original, forgotten spelling, as it could be seen in not so old and common dictionaries like *Petit Larousse* (see below).
There was clearly a gap to be filled, in order to bring to light recent evolutions of the language.



From 1955

1906-1947 **NÉNUFAR** ou **NÉNUPHAR** n. m. Genre de nymphéacées aquatiques, à larges feuilles et à fleurs jaunes ou blanches, qui croissent dans les pays chauds et tempérés : *le nénufar blanc est le lotus sacré des Égyptiens.*

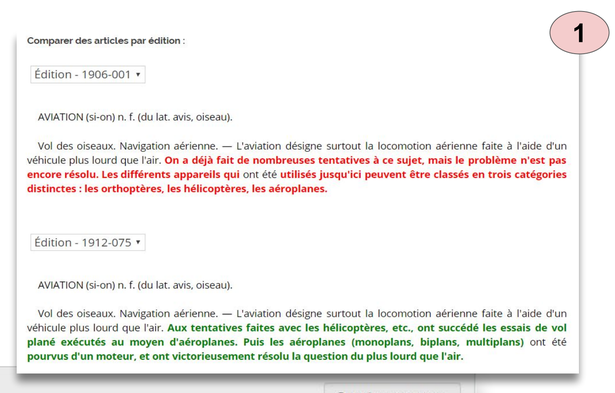
1948-1954 **NÉNUPHAR** ou **NÉNUFAR** n. m. Genre de nymphéacées aquatiques, à larges feuilles et à fleurs jaunes ou blanches, qui croissent dans les pays chauds et tempérés : *le nénuphar blanc est le lotus sacré des Égyptiens.*

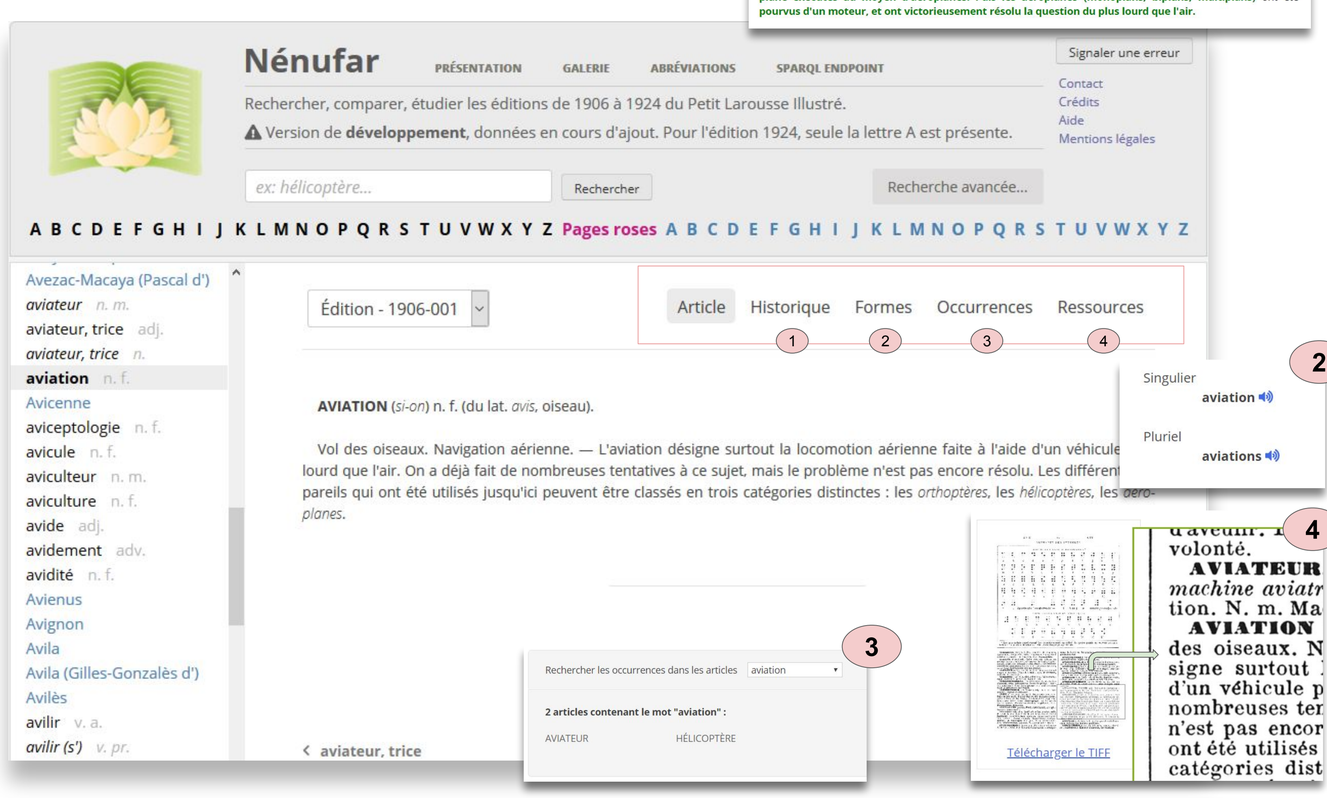
NÉNUPHAR n. m. Genre de nymphéacées aquatiques, à larges feuilles et à fleurs jaunes ou blanches, qui croissent dans les pays chauds et tempérés : *Le nénuphar blanc est le lotus sacré des Égyptiens.*

Some word datations:
1908: autobus, gréviste (*striker*)
1911: antimilitarisme
1912: puériculture, technicien
1914: tract
1915: copyright
1916: scientisme, syndicalisme
1918: boche *(pej. for German)*
1920: alerter
1921: bolchévik, bolchévisme
1922: mazout, pogrom, soviet

A new resource for studying recent history of language, culture and techniques

An example with the entry **AVIATION** between 1906 and 1912 editions





WEB site

Targeting both scientific (linguists, historians, lexicographers, social scientists) and general public. Simple and advanced research criteria, user friendly interface, responsive design.
Public release in July 2018 with 1906 to 1924 editions, then new data will be regularly added.

TEI XML

Reference digital edition format. For long-term preservation and expert querying.
Users: digital humanists, computational linguists.
The use of TEI-LEX0 format (Banski *et al.*, 2017) will allow for a predictable structure and for the exploitation of *ad hoc* digitisation technologies such as GROBID-Dictionaries (Khemakhem *et al.,* 2017)

Ontolex-Lemon RDF

Natural language processing applications and linking to other resources.
Users: language technologists, computer scientists.

```
<entry xml:id='aviation' n='1906-001_1912-075'>
    <form type='lemma'>
        <orth>AVIATION</orth>
        <pron>(<emph rend='italic'>si-on</emph>)</pron>
    </form>
    <gramGrp><pos>n.</pos> <gen>f.</gen></gramGrp>
    <etym>(du <lang expand='latin'>lat.</lang> <mentioned>avis</mentioned>, <gloss>oiseau</gloss>).</etym>
    <sense>
        <def>Vol des oiseaux.</def>
    </sense>
    <sense>
        <def>Navigation aérienne.</def>
        <note type='encycl'>— L'aviation désigne surtout la locomotion aérienne faite à l'aide d'un véhicule plus
            lourd que l'air. On a déjà fait de nombreuses tentatives à ce sujet, mais le problème n'est pas encore
            résolu. Les différents appareils qui ont été utilisés jusqu'ici peuvent être classés en trois catégories
            distinctes : les <mentioned>orthoptères</mentioned>, les <mentioned>hélicoptères</mentioned>, les
            <mentioned>aéroplanes</mentioned>.
        </note>
    </sense>
</entry>
```

4 Afficher l'article au format XML

References:
Banski, P., Bowers, J., and Erjavec, T. (2017). TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms. In eLex 2017.
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.
Bohbot, H., Frontini, F., Luxardo, G., Khemakhem, M., & Romary, L. (2018). Presenting the Nénufar Project: a Diachronic Digital Edition of the Petit Larousse Illustré. In GLOBALEX 2018 - Globalex workshop at LREC2018 (pp. 1–6). Miyazaki, Japan.
Pruvost, J. (2002), *Les dictionnaires de langue française*. Coll. Que-sais-je ? n°3622, PUF Paris.

Info: projet-nenufar@univ-montp3.fr

Nénufar project is headed by laboratoire **Praxiling** at Paul-Valéry University of Montpellier in collaboration with **Inria**. It is supported by the *Délégation Générale à la Langue Française et aux Langues de France* (an agency of French Culture and Communication Ministry) and the Huma-Num consortia **CORLI** and **CAHIER**.