Nénufar: Modelling a Diachronic Collection of Dictionary Editions as a Computational Lexical Resource

Hervé Bohbot[1], Francesca Frontini[1], Fahad Khan[2]

Mohamed Khemakhem[3,4,5] and Laurent Romary[3,4,6]

[1] Praxiling UMR 5267 CNRS, Université Paul-Valéry Montpellier 3
[2] Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa
[3] ALMAnaCH - INRIA
[4] CMB - Centre Marc Bloch, Berlin
[5] UPD7 - Université Paris Diderot, Paris 7
[6] BBAW - Berlin-Brandenburgische Akademie der Wissenschaften, Berlin

E-mail: herve.bohbot@cnrs.fr, francesca.frontini@univ-montp3.fr, fahad.khan@ilc.cnr.it, mohamed.khemakhem@inria.fr, laurent.romary@inria.fr

Abstract

The *Petit Larousse Illustré* (PLI) is a monolingual French dictionary which has been published every year since the 1906 edition[1] and **WHICH** is therefore a fundamental testimony of the evolution of the French language. As a consequence of the pre-1948 editions of the PLI entering the public domain in 2018 the *Nénufar* ("Nouvelle édition numérique de fac-similés de reference") project was launched at the Praxiling laboratory in Montpellier with the aim of digitising and make these editions available electronically. The project is still ongoing; various selected editions per decade are going to be fully digitised (so far the 1906, 1924 and 1925 editions have been completed), and changes backtracked and dated per specific year.

Nénufar's primary aim is to make the editions available and searchable via an advanced search interface which will not only enable the selective querying of text by lemma and type of content (definitions, examples, ...), but crucially also detect and study changes by comparing different editions. In order to do so, a specific web interface has been put in place (see Figure 1 and the project's website[2]). Alongside the digitised text, the Nénufar web site[3] contains high quality scans for each page. In

[1] Published in 1905 but dated 1906.

[2] http://nenufar.huma-num.fr/?article=3807

[3] A similar project, which presents data and scans from subsequent editions of the same legacy dictionary has been carried out by the team behind the Swedish Academy's Wordlist (see Holmer, Malmgren & Martens, 2016 and http://spraakdata.gu.se/saolhist/).

compliance with current open data best practices (Wilkinson et al., 2016), the project also aims to make the source data available separately from the querying interface both for research and for long term preservation. The primary encoding format is TEI-XML; however in our case the TEI encoding is closely inspired from the latest version of the TEI-Lex0 (Bański et al., 2017, Romary & Tasovac, 2018) guidelines for encoding lexicographic resources[4], which are based upon TEI. The choice of a TEI[5] based approach allows the Nénufar project to align itself to other pre-existing initiatives and tools. By aligning ourselves to TEI-Lex0 we will be able to make use of digitisation tools such as *Grobid* (Khemakhem et al., 2017) which have TEI-Lex0 as their native format and which have already been tested and used within the Nénufar project to speed up the digitisation of new editions. In addition we will be able to make use of ongoing initiatives to convert TEI-Lex0 datasets to RDF using the W3C recommendation for publishing lexicons as Linked Data, namely Ontolex Lemon (McCrae et al., 2017; Bosque-Gil et al., 2016) which will allow for the publication of the Nénufar dataset as a LOD graph. The LOD version of the Nénufar dataset, now currently being developed, will be queryable from the available SPARQL endpoint and contain all available editions as one single graph allowing for expert users to perform complex queries that could detect systematic changes in the dataset. The LOD version is particularly adapted to be linked to other datasets; more recent editions, once added, could also be of interest for NLP applications.

The presentation will illustrate the state of the art of the project, showcase the web site, outline the principles guiding the TEI encoding with examples, and discuss the issues concerning the conversion from TEI-Lex0 to Ontolex Lemon.

References

Bański, P., Bowers, J. and Erjavec, T. (2017). TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms. *ELex2017*.

Bohbot, H., Frontini, F., Luxardo, G., Khemakhem, M. and Romary, L. (2018). Presenting the Nénufar Project: a Diachronic Digital Edition of the Petit Larousse Illustré. *GLOBALEX 2018 - Globalex Workshop at LREC2018*. Miyazaki, Japan.

Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E. and Aguado-de-Cea, G. (2016). Modelling Multilingual Lexicographic Resources for the Web of Data: the K Dictionaries case. *GLOBALEX 2018 - Globalex Workshop at LREC2018*. Miyazaki, Japan.

Holmer, Louise, Sven-Göran Malmgren, and Monica von Martens. (2016).

[4] We chose not to encode the dataset directly into TEI-Lex0 since the guidelines are still incomplete. But our dataset is as closely aligned as possible to the current version of the TEI-Lex0 guidelines in order to make any future conversion as straightforward as possible.

[5] On the Nénufar website select an entry and then “Ressources” to inspect the xml encoding.

"SAOLhist.se – för allmänt och vetenskapligt bruk." *Nordiske Studier i Leksikografi*, no. 13: 349–58.

Khemakhem, M., Foppiano, L. and Romary, L. (2017). Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. *Electronic Lexicography, ELex 2017*. Leiden, Netherlands.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P. and Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. *ELex2017*.

Romary, Laurent, and Toma Tasovac, 'TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources' (presented at the TEI Conference, Tokyo, 2018) <https://hal.inria.fr/hal-02265312>

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3: 160018 doi:10.1038/sdata.2016.18.

Figure 1: Comparing the differences in the definition of "aviation" (aviation) between 1906 and 1912 edition. In the first one air navigation is still a possibility, whereas in the second one it is a reality.