

Preparing the Dictionnaire Universel for Automatic Enrichment

Pedro Javier ORTIZ SUAREZ^{1,2}, Laurent Romary¹ Benoît Sagot¹ June 12, 2019

¹Inria, Paris, France ²Sorbonne Université, Paris, France

- 1. Introduction
- 2. Automatic Enrichment of Ancient Dictionaries
- 3. Workflow
- 4. Preparing the Dictionnaire Universel for Automatic Enrichment

Introduction

The Dictionnaire Universel

- In 1635, the Académie française began producing a dictionary of the French language [2]. Antoine Furetière got involved in the production of the first edition.
- Furetière grew frustrated with the slow progress of the Académie. He began to work on his own dictionary, The Dictionaire universel (DU) around 1676-78 [7].
- The second edition (1701) of the *DU* was improved by *Henri Basnage de Beauval* who virtually turned the *DU* into an encyclopaedia [1][3].



The aim of **BasNum** is to digitize the DU, in the 1701 version rewritten by Basnage de Beauval, to analyse it with computational methods in order to better assess the importance of this work for the evolution of sciences and mentalities in the 18th century, and to contribute to the contemporary movement for creating innovative and data-driven computational methods for text digitization, encoding and analysis.



Automatic Enrichment of Ancient Dictionaries

At *Inria* we will develop methods for the automatic structuring of lexicographical entries from the digitized version of Basnage's *DU*, we will also enrich the contents of this dictionary using machine learning methods. There are two main types of machine learning models:

• Statistical methods: We have some existing tools.

• Neural methods: We want to build our own tools.



Optical Character Recognition

Transkribus

- Well established tool.
- Uses statistical methods.
- Has an user interface.
- Generates PDF's.
- Does not preserve typographical features.
- Free but not Open Source.
- Transkribus infrastructure has to be used.

Kraken

- Not so well established.
- Uses neural methods.
- User interface in development.
- Does not generate PDF's.
- Preserves Typographycal features.
- Free and Open Source.
- We can use our own infrastructure.

GROBID-Dictionaries [4][5] is an machine learning library using statistical methods for structuring digitised lexical resources and entry-based documents with encyclopedic or bibliographic content. It allows the parsing, extraction and structuring of text information in such resources.

GROBID-Dictionaries takes as input lexical resources digitised in PDF format and generates a TEI-encoded hierarchy of the different recognised text structures.

GROBID-Dictionaries

- RAQUETTE. f.f. Espece de pallette pour jouër à la paume, & au volant. Elle est faite d'un treillis de cordes de boyaux (dont les unes s'appellent montans, & les autres travers) fort tenduës sur nour de bois qui a un manche de mediocre longueur. Un de ses côtez s'appelle les droits, & l'autre les næuds. Pasquier a remarqué qu'anciennement on ne joüoit point à la paume avec des raquettes: c'étoit avec la paume de la main; & de là il conjecture qu'est venu le nom de jeu de pamme. On n'avoit inventé les raquettes qu'un peu avant le temps de Pasquier, à ce qu'il dit.
- Menage derive ce mot de retiquetta, diminutif de retis, reticus & reticulum.
- On dit proverbialement pour fe mocquer d'un homme qui fe vante de plufieurs chofes qu'il n'a pas faites, C'eft un grand catieur de *raquettes*.
- RAQUETTE, se dit aussi d'une certaine machine que les Sauvages de Canada attachent à leurs pieds pour marcher plus commodément sur la neige, & qui est faite à-peu-près en forme de *raquette* à jouër.
- RAQUETTE, fe dit auffi d'une espece de figuier d'Inde qui croît aux lles Antilles: c'est cette espece que Mr. Tournesort appelle opuntia vulgo herbariorum, J. BAUH. Ses seuilles sont épaisses, longues, quelquesois larges comme une raquette, d'où vient que les François lui ont donné ce nom. Voyez FIGUIER D'INDE.



GROBID-Dictionaries

<div n="RAQ">

cetym-bibl>cauthor ref="#Men_ISM00000008885971">Menoge</author=</br/>obibl> derive ce mot de <lang rendition="#i" wrlichag="lat">retiqueta</lang>, diminutif de retis, reticus & reticulum.</retym>

<sense-xusg>On dit proverbialement</usg> <gloss-pour se mocquer d'un homme qui se vante de plusieurs choses qu'il n'a pas faites,</gloss> <seg type="Proverbe">C'est un grand casseur de raquettes,</sense>

<sense-form-orth rend(tion="\$c5">Saquette_/orth>-/form-<def>se dit aussi d'une certaine machine que les Sauvages de Canada attenta à leurs jeds pour marcher plus commodément sur la neige, samp; qui est faite à peu-près en forme de raquette à joufe-//def/sense>

-sense-form-orth rend(tion="ssc"-Raquete, /orth-/form-/def/se dit aussi d'une espece de figuier d'Inde qui croît aux lles Artilles ://def, =notec'est cette espece que Mr. Tournefort appelle opuntia vule herbariorum, J. Bauh. Ses fauilles sont épaisses, longues, quelquefois larges comme une raquette, d'aù vient que les François lui ont domé com./notes -cmr.Voyez -ref corresp="#ficuier">-Ficuier d'Inde./refs/sense>

</div>









Deep Learning Model





Crowdsourcing Model



Preparing the Dictionnaire Universel for Automatic Enrichment

Recommendations

• Annotate intelligently for not so intelligent models! At the beginning, only entries for the letter C where transcribed. Our OCR engine learned that every entry starts with C. Randomise your annotations!

\cdot TEI is flexible

TEI is flexible to some extend, **agree** on how to annotate, as multiple annotations schemes reduce the precision of the model.

• Get a lot of text in the target language [6] Then get a lot of data of the target period or about the target topic.

Recommendations

• Try to get a high quality data to reduce noise.

DICTIONNAIRE UNIVERSEL

Contenant generalement tous les

MOTS FRANCOIS

tant vieux que modernes, & les Termes des

SCIENCES ET DES ARTS.

A.

A THE AT con au étrieur coffenere , se support lantes com qui étriren excluente , ne manques janvis d'ablerver cone différence. Constante a étrie sufficience.

rale qui lignificit 500, comune on voit dans Velerins Frohm, il y a des von meiers rappentes per Barmina, qui marquere les hurses fignificatives des nombres, dans la recenter effe

don in pienter ch: Teller A server quiegets sofar rela. Quest or sensitiv sofar occur ligor docht zu deflar de 14. Hagefals eine anfle. Les Beneine Tappf-biern Ante Quester, servergeren fen ferreite pour dechters Factoff insorten. A vositik deradjöfer, je Publicas, Carso large a diverse femifications. Creendant il en faut

Cavilians a distribut fighteeness. Copendate II mfore Views In second responses and associated and the Views II second response and the distance Cult application in a distribution of the distance Cult application in a distribution of the distance of the application in a distribution of the distribu-tion for the provide the second and the distribu-tion of the application of the distribution of the second and the distribution of the distribution of the distribution of the distribution of the Cult of provises in the Ministry of the distribution Cult of the distribution of the distribution of the Cult of the distribution of the distribution of the Cult of the distribution of the Cult of the distribution of the distribution of the distribution of the Cult of the distribution of

a fair de Péche sual à proper. L'imagentien du Poine n'a pa vous pénder fabrile que vous fari, Vou. La versió qui a des bonnes, a dit pour vous tout ce que la

Langung payer distribution of exhibiting in early of the second part of the second par (a) A set of the se

grave, prevedues here on transitionmemory of an issues. The formation of the symbol memory beneric series of the symbol memory of the bin adversity. Lapping of the symbol memory restored a symbol memory of the symbol me

DICTIONNAIRE UNIVERSEL.

Contenant generalement tous les

MOTS FRANCOIS

tant vieux que modernes', & les Termes des

SCIENCES ET DES ARTS

A. A. Premiere lettre de l'Alphaber François, Els time le mêsse raçe dans source les largest constain, accord la larges F des della polat une royole, Cell are

predimension in Program and State an e prensier en & le prensier begsymmene des enfans. Il and region pose cont, man tant i entertariany de cos vos-gets, en pagen qui est long, quand il fera nurqui d'an perida i construct, et bori quand il seura pos eracerat, car A ell formen un anche pour declarer les rooma approve

qui masquent les leures farificatives des nombres,

Qued ce esettoi un tion ou une ligne droire su deffar de l'A, Afgustore cuq nafle. Les Romais l'appel-laires lorre feinere, parcept'on s'un farrois pour feelure l'accifé instrute. A vonkit dire alfabe, je

Concloure a loverfes familications, Cependare il en fast rade & moins agreable. Ceft qualqueries un fabilimit mifenin. Cet A eft mal

• A fait de l'éche mul à-proper. L'imagination du Proce-né pa vous pender fichelle que nou den. Vo., La vertifiq qui à de hornes, a du procevon trace qui la mendege qui d'accemb point, a traverte fore la mo-nes. G. M. Dans cerne tignifications l'on s'y met point d'accemb, si quard il el proced à la particle y, par dons un la forcada voite foldenti des. Il y a par dons un la forcada voite foldenti des. Il y a

Melacitel Quad on 6 fere affordir: He je me menni Dan h correlation: Ab? Moniers, por ce vers je von denandegreet. Box. Dans l'écontencer: Ab

Ceffand Hall, poor experience movies.
Gelgen gi be, Lobi, Aller, and Laborand Hanger, Jane Sagestammerganisk downer etersan tor mangeer poor Sagestammerganisk downerganisk downerganisk downerganisk downerganisk downergan

Thank you!

References i

 H. Bots, L. Van Lieshout, and H. B. de Beauval.
Contribution à la connaissance des réseaux d'information au début du XVIIIe siècle: Henri Basnage de Beauval et sa correspondance à propos de l'" Histoire des ouvrages des savans", 1687-1709: publication annotée de quelque cent lettres et index thématique et analytique, volume 3.
Amsterdam, Holland University Press, 1984.

J. Considine.

Academy Dictionaries 1600–1800.

Cambridge University Press, 2014.

References ii

A. Furetière.

Dictionnaire Universel, contenant généralement tous les mots françois tant vieux que modernes, & les termes des sciences et des arts.

Arnoud et Reinier Leers, 1701.

Par Feu Messire Antoine Furetière,... 2e édition revue, corrigée et augmentée par M. Basnage de Bauval.

M. Khemakhem, L. Foppiano, and L. Romary. Automatic extraction of tei structures in digitized lexical resources using conditional random fields. electronic lexicography, eLex 2017, Sept. 2017.

References iii

M. Khemakhem, A. Herold, and L. Romary. Enhancing usability for automatically structuring digitised dictionaries.

GLOBALEX workshop at LREC 2018, May 2018.

 P. J. Ortiz Suárez, B. Sagot, and L. Romary.
Asynchronous Pipelines for Processing Huge Corpora on Medium to Low Resource Infrastructures.

In 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7), Cardiff, United Kingdom, July 2019.

A. Rey.

Antoine Furetière, imagier de la culture classique, Le dictionnaire universel d'Antoine Furetière, volume 1. Paris: SNL - Le Robert, 1978.