Preparing the Dictionnaire Universel for Automatic Enrichment

Pedro Javier Ortiz Suárez, Laurent Romary, Benoît Sagot

Preparing the Dictionnaire Universel for Automatic Enrichment

Pedro Javier Ortiz Suarez[1,2], Laurent Romary[1] Benoît Sagot[1]

June 12, 2019

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

Table of contents

Abstract

Abstract

The Dictionnaire Universel (DU) is an encyclopaedic dictionary originally written by Antoine Furetière around 1676-78, later revised and improved by the Protestant jurist Henri Basnage de Beauval who expanded, corrected and included terms of arts, crafts and sciences, into the Dictionnaire. The aim of the BASNUM project is to digitize the DU in its second edition 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. Based on the experience acquired within the research group, an enrichment workflow based upon a series of Natural Language Processing processes is being set up to be applied to Basnage's work. This includes, among others, automatic identification of the dictionary structure (macro-, meso- and microstructure), named-entity recognition (in particular persons and locations), classification of dictionary entries, detection and study of polysemy markers, tracking and classification of quotation use (bibliographic references), scoring semantic similarity between the DU and other dictionaries. The main challenges being the lack of available annotated data in order to train machine learning models, decreased accuracy when using modern pre-trained models due to the differences between present-day and 18th century French, and even unreliable or low quality OCRisation. The paper describes methods that are useful to tackle these issues in order to prepare the the DU for automatic enrichment going beyond what current available tools like Grobid-dictionaries can do, thanks to the advent of deep learning NLP models. The paper also describes how these methods could be applied to other dictionaries or even other types of ancient texts.

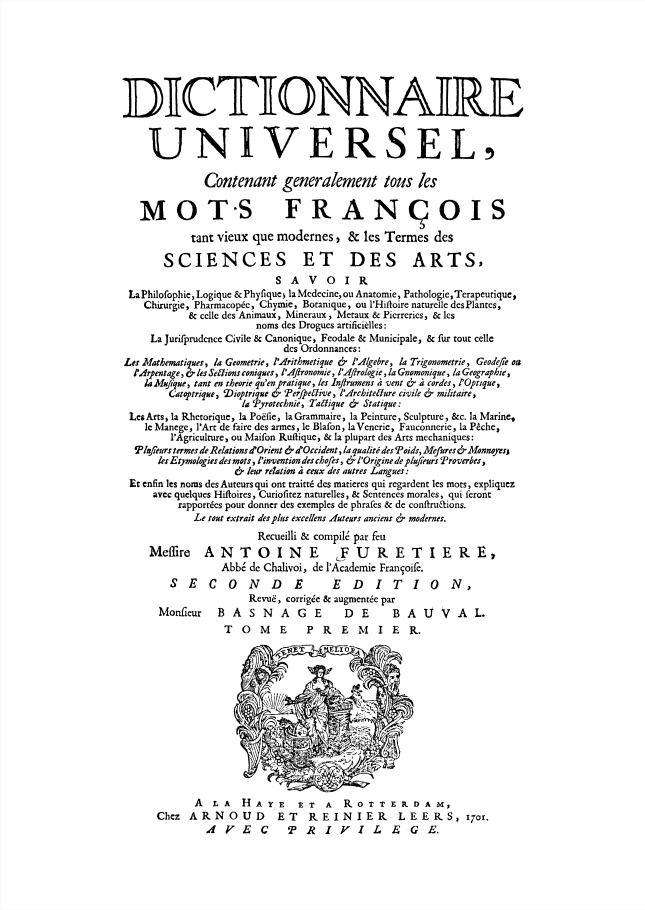
Introduction

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 BasNum Project

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.



DICTIONNAIRE
UNIVERSEL,
Contenant generalement tous les
MOTS FRANÇOIS
tant vieux que modernes, & les Termes des
SCIENCES ET DES ARTS.
SAVOIR
La Philoſophie, Logique & Phyſique, la Medecine, ou Anatomie, Pathologie, Terapeutique, Chirurgie, Pharmacopée, Chymie, Botanique, ou l'Hiſtoire naturelle des Plantes, & celle des Animaux, Mineraux, Metaux & Pierreries, & les noms des Drogues artificielles:
La Juriſprudence Civile & Canonique, Feodale & Municipale, & ſur tout celle des Ordonnances:
Les Mathematiques, la Geometrie, l'Arithmetique & l'Algebre, la Trigonometrie, Geodeſie ou l'Arpentage, & les Sections coniques, l'Aſtronomie, l'Aſtrologie, la Gnomonique, la Geographie, la Muſique, tant en theorie qu'en pratique, les Inſtruments à vent & à cordes, l'Optique, Catoptrique, Dioptrique & Perſpective, l'Architecture civile & militaire, la Pyrotechnie, Tactique & Statique:
Les Arts, la Rhetorique, la Poëſie, la Grammaire, la Peinture, Sculpture, &c. la Marine, le Manege, l'Art de faire des armes, le Blaſon, la Venerie, Fauconnerie, la Pêche, l'Agriculture, ou Maiſon Ruſtique, & la plupart des Arts mechaniques:
Pluſieurs termes de Relations d'Orient & d'Occident, la qualité des Poids, Meſures & Monnoyes, les Etymologies des mots, l'invention des choſes, & l'Origine de pluſieurs Proverbes, & leur relation à ceux des autres Langues:
Et enfin les noms des Auteurs qui ont traitté des matieres qui regardent les mots, expliquez avec quelques Hiſtoires, Curioſitez naturelles, & Sentences morales, qui ſeront rapportées pour donner des exemples de phraſes & de conſtructions.
Le tout extrait des plus excellens Auteurs anciens & modernes.
Recueilli & compilé par feu
Meſſire ANTOINE FURETIERE,
Abbé de Chalivoi, de l'Academie Françoiſe.
SECONDE EDITION,
Revuë, corrigée & augmentée par
Monſieur BASNAGE DE BAUVAL.
TOME PREMIER.
A LA HAYE ET A ROTTERDAM,
Chez ARNOUD ET REINIER LEERS, 1701.
AVEC PRIVILEGE.

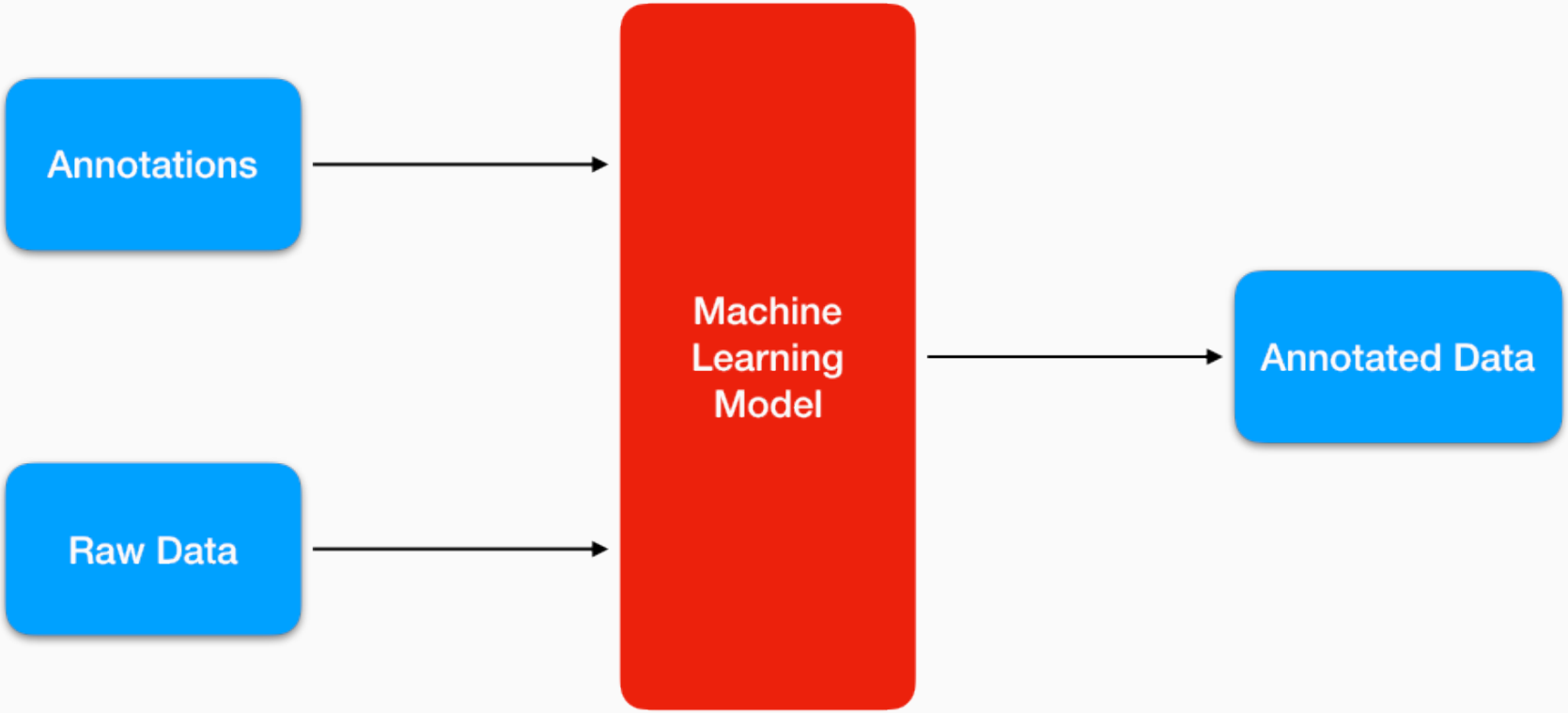
Automatic Enrichment of Ancient Dictionaries

BasNum at Inria

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.

Machine Learning



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

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. ſ. f. Eſpece de pallette pour joüer 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 un tour de bois qui a un manche de mediocre longueur. Un de ſes côtez s'appelle les *droits*, & l'autre les *nœuds*. 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 *paume*. On n'avoit inventé les *raquettes* qu'un peu avant le temps de Paſquier, à ce qu'il dit.

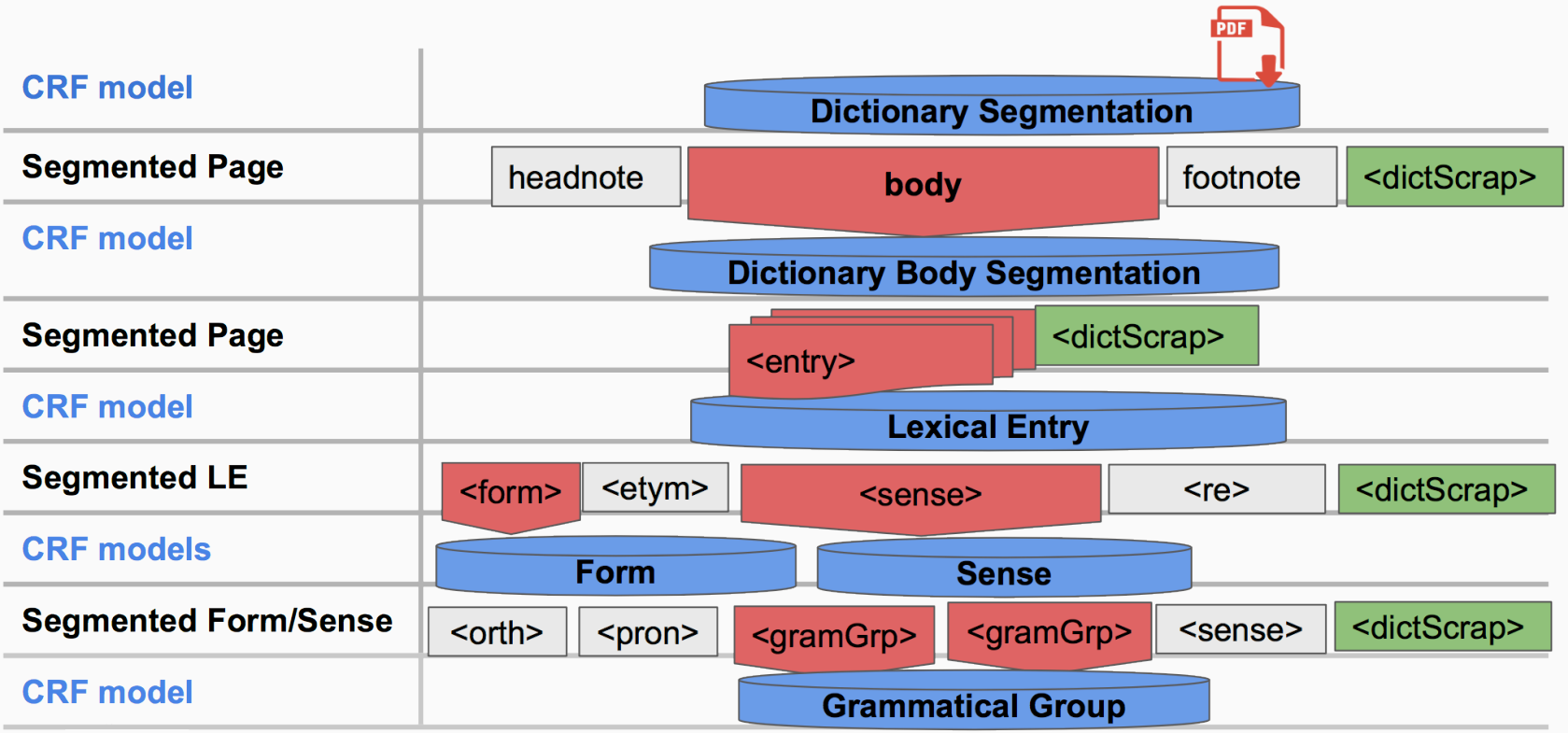
Menage derive ce mot de *retiquetta*, diminutif de *retis*, *reticus* & *reticulum*.

On dit proverbialement pour ſe mocquer d'un homme qui ſe vante de pluſieurs choſes qu'il n'a pas faites, C'eſt un grand caſſeur 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 joüer.

RAQUETTE, ſe dit auſſi d'une eſpece de figuier d'Inde qui croît aux Iles Antilles: c'eſt cette eſpece que Mr. Tournefort appelle *opuntia vulgò herbariorum*, J. BAUH. Ses feuilles ſont épaiſſes, longues, quelquefois larges comme une raquette, d'où vient que les François lui ont donné ce nom. Voyez FIGUIER D'INDE.

GROBID-Dictionaries

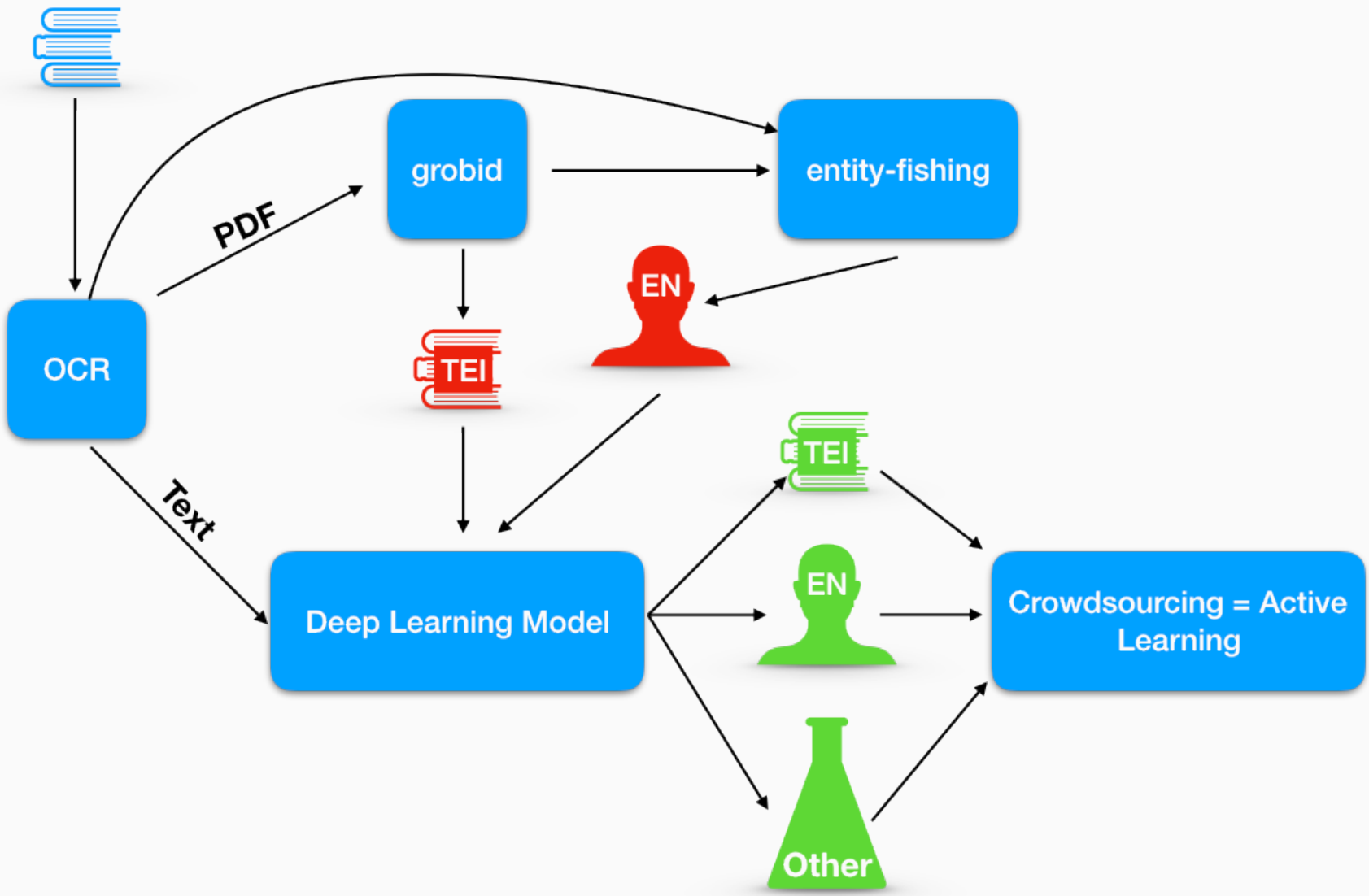


GROBID-Dictionaries

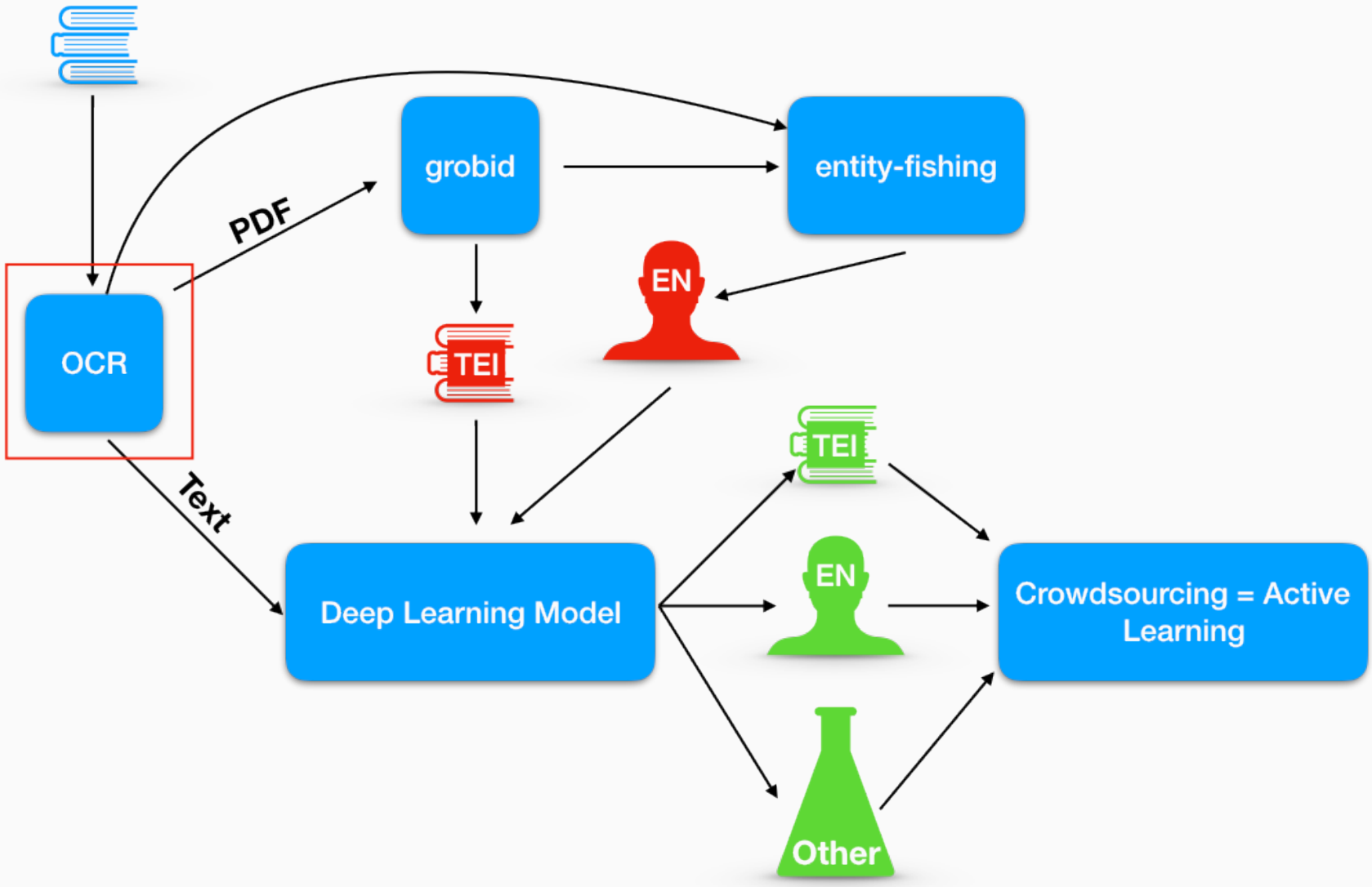
```
<div n="RAQ">
    <entry xml:lang="fre" xml:id="Raquette">
        <sense><form><orth rendition="#uc">raquette.</orth><gramGrp><pos expand="Substantif">s.</pos>
        <gen expand="Feminin">f.</gen></gramGrp></form>  <def>Especе de pallette pour jouër
        à la paume, &amp;amp; au volant.</def> <note>Elle est faite d'iun treillis de
        cordes de boyaux édont les unes s'appellent montans
        &amp;amp; les autres travers J fort tendues sur un tour de bois
        qui a un manche de mediocre longueur.</note> <xr corresp="#droits #noeuds">Un de ses cô-
        tea s'appelle les droits, &amp;amp; l'autre les nœuds</xr>. <etym><bibl><author ref="#Pasquier_ISN000000010907010">Pasquier</author></bibl>
        a remarqué qu'anciennement on ne joüoit point à la
        paume avec des raquettes : c'étoit avec la paume de la
        main , &amp;amp; de là il conjectue qu'est venu le nom de jeu
        de pame.</etym> <note>On n'avoit inventé les raquettes qu'un peu
        avant le temps de <bibl><author ref="#Pasquier_ISN000000010907010">Pasquier</author></bibl>, à ce qu'il dit.</note>


        <etym><bibl><author ref="#Men_ISN0000000080815971">Menage</author></bibl> derive ce mot de
        <lang rendition="#i" xml:lang="lat">retiquetta</lang>, diminutif de retis,
        reticus &amp;amp; reticulum.</etym>
        </sense>

        <sense><usg>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.</seg></sense>


        <sense><form><orth rendition="#sc">Raquette,</orth></form> <def>se dit aussi d'une certaine machine que
        les Sauvages de Canada attachent à leurs pieds pour
        marcher plus commodément sur la neige, &amp;amp; qui est
        faite à-peu-près en forme de raquette à jouër.</def></sense>


        <sense><form><orth rendition="#sc">Raquette,</orth></form> <def>se dit aussi d'une espece de figuier d'In-
        de qui croît aux lles Antilles :</def> <note>c'est cette espece que
        Mr. Tournefort appelle opuntia vule herbariorum,
        J. Bauh. Ses feuilles sont épaisses, longues, quel-
        quefois larges comme une raquette, d'où vient que les
        François lui ont donné ce nom.</note> <xr>Voyez <ref corresp="#Ficuier">Ficuier
        d'Inde.</ref></xr></sense>
    </entry>
</div>
```

Workflow

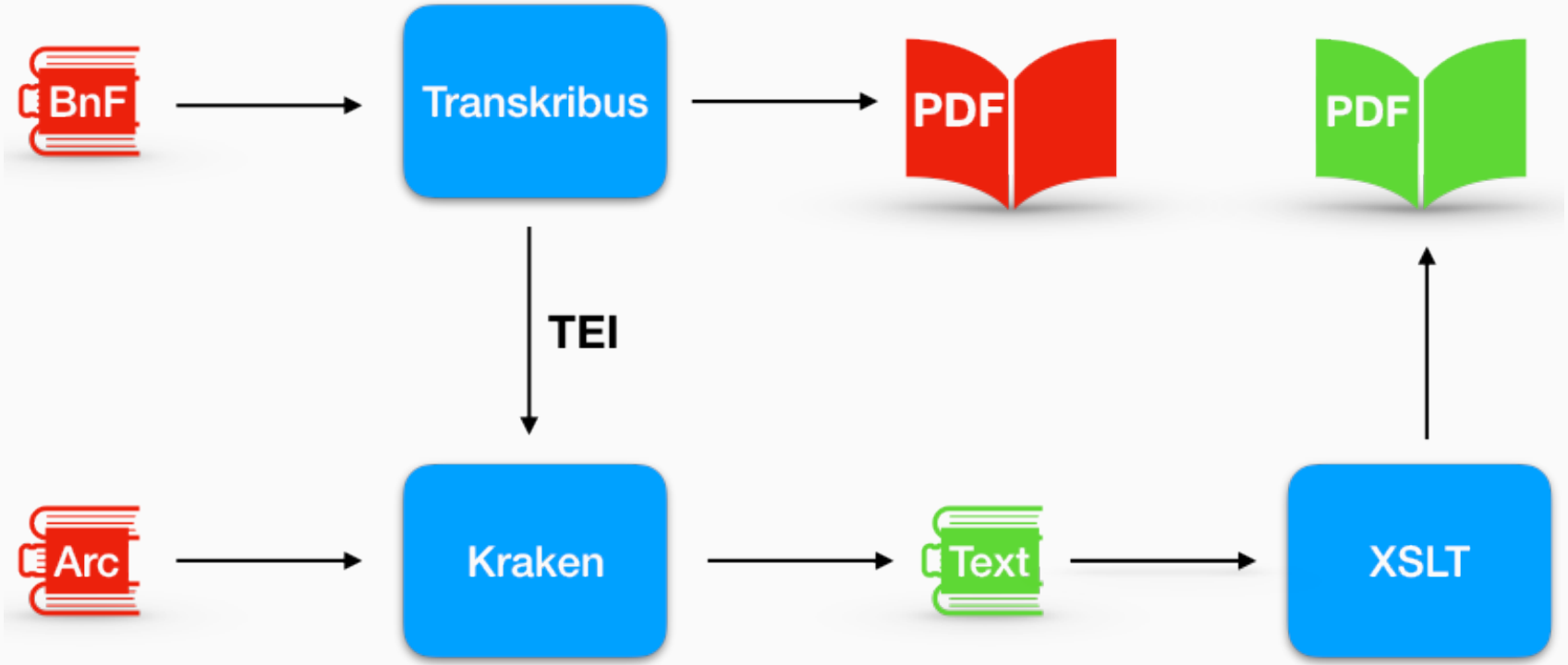
Workflow


grobid
entity-fishing
PDF
OCR
TEI
EN
Text
TEI
Deep Learning Model
EN
Crowdsourcing = Active Learning
Other

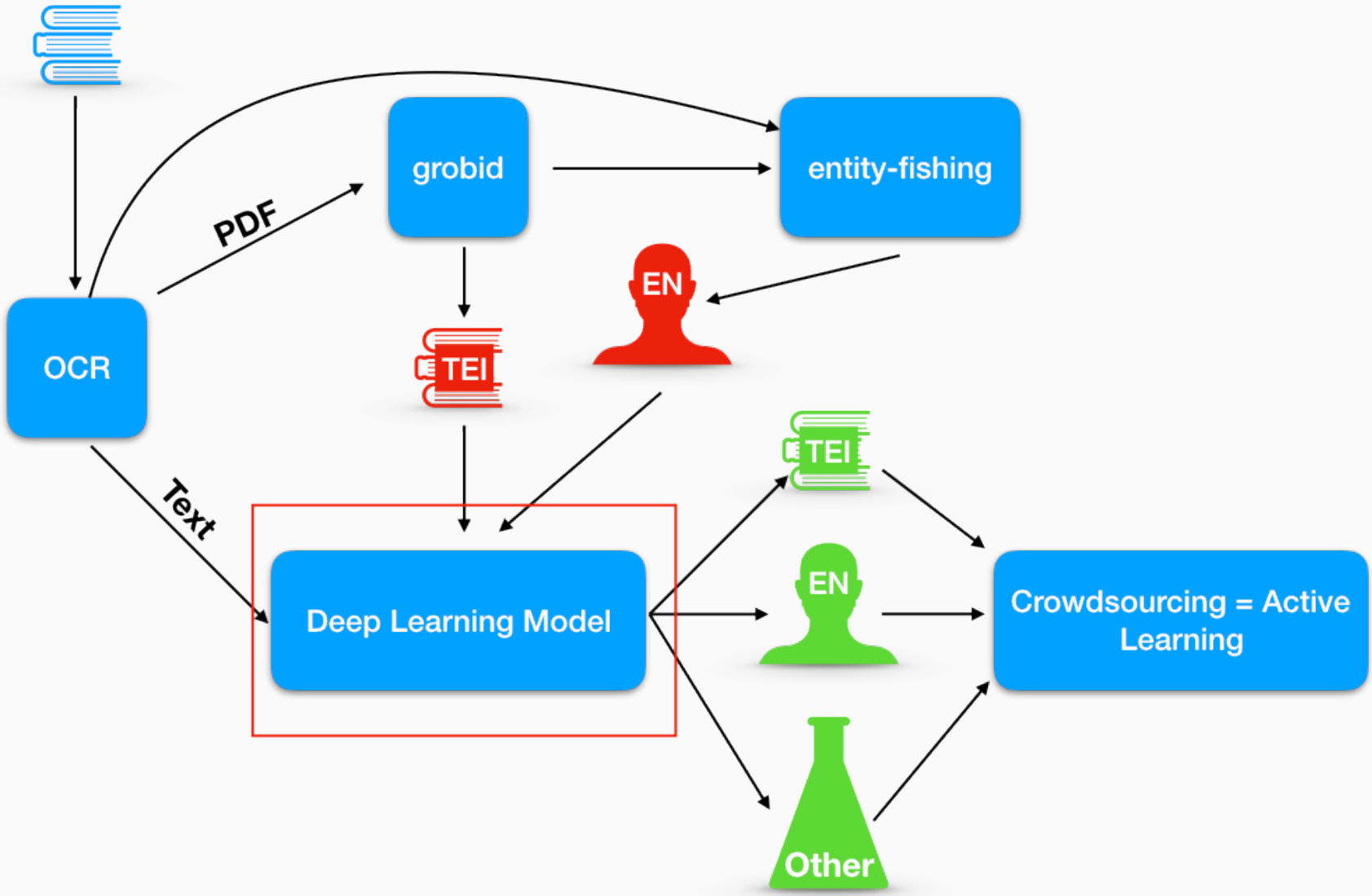
Workflow


grobid
entity-fishing
PDF
OCR
TEI
EN
Text
TEI
Deep Learning Model
EN
Crowdsourcing = Active Learning
Other

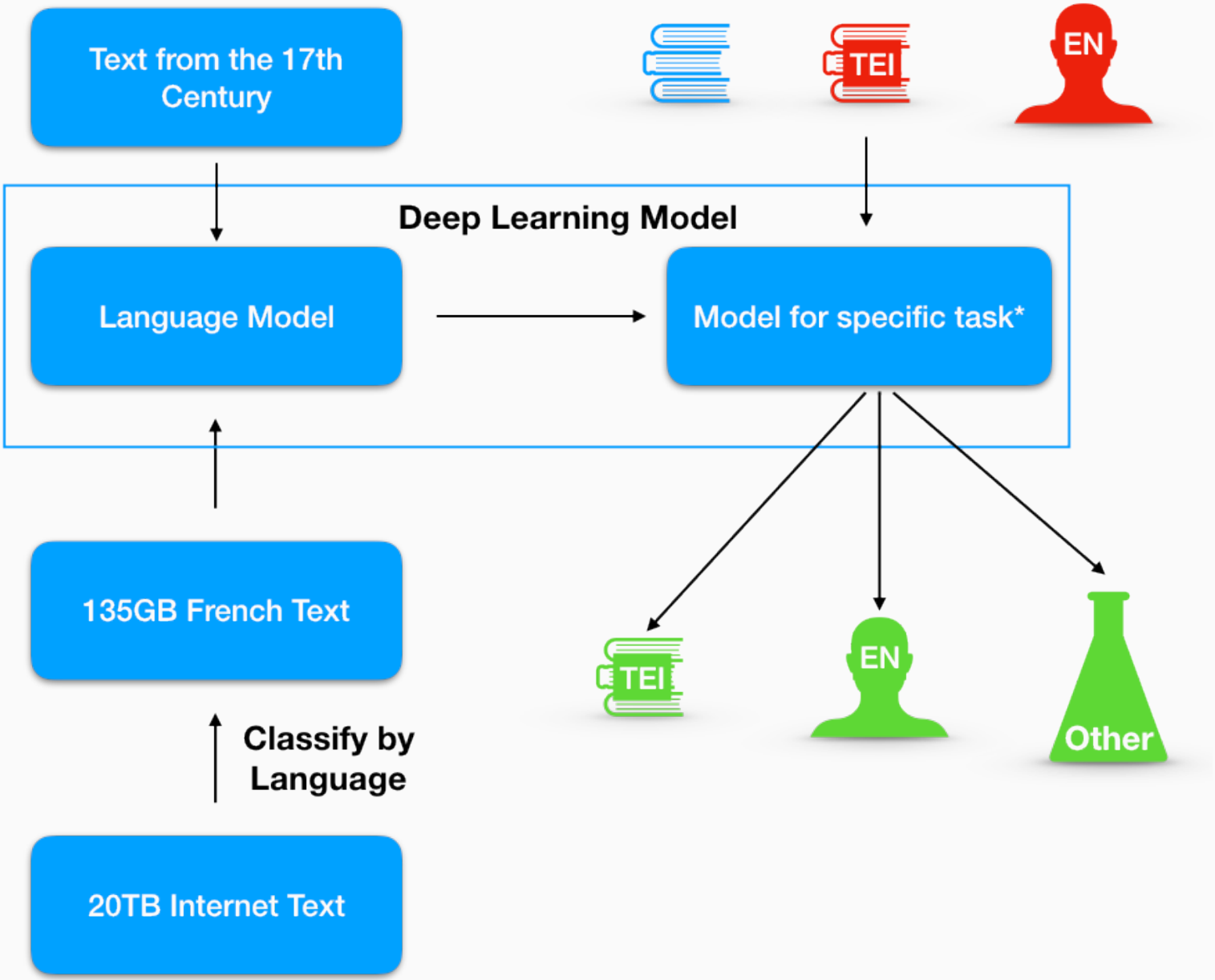
OCR Model



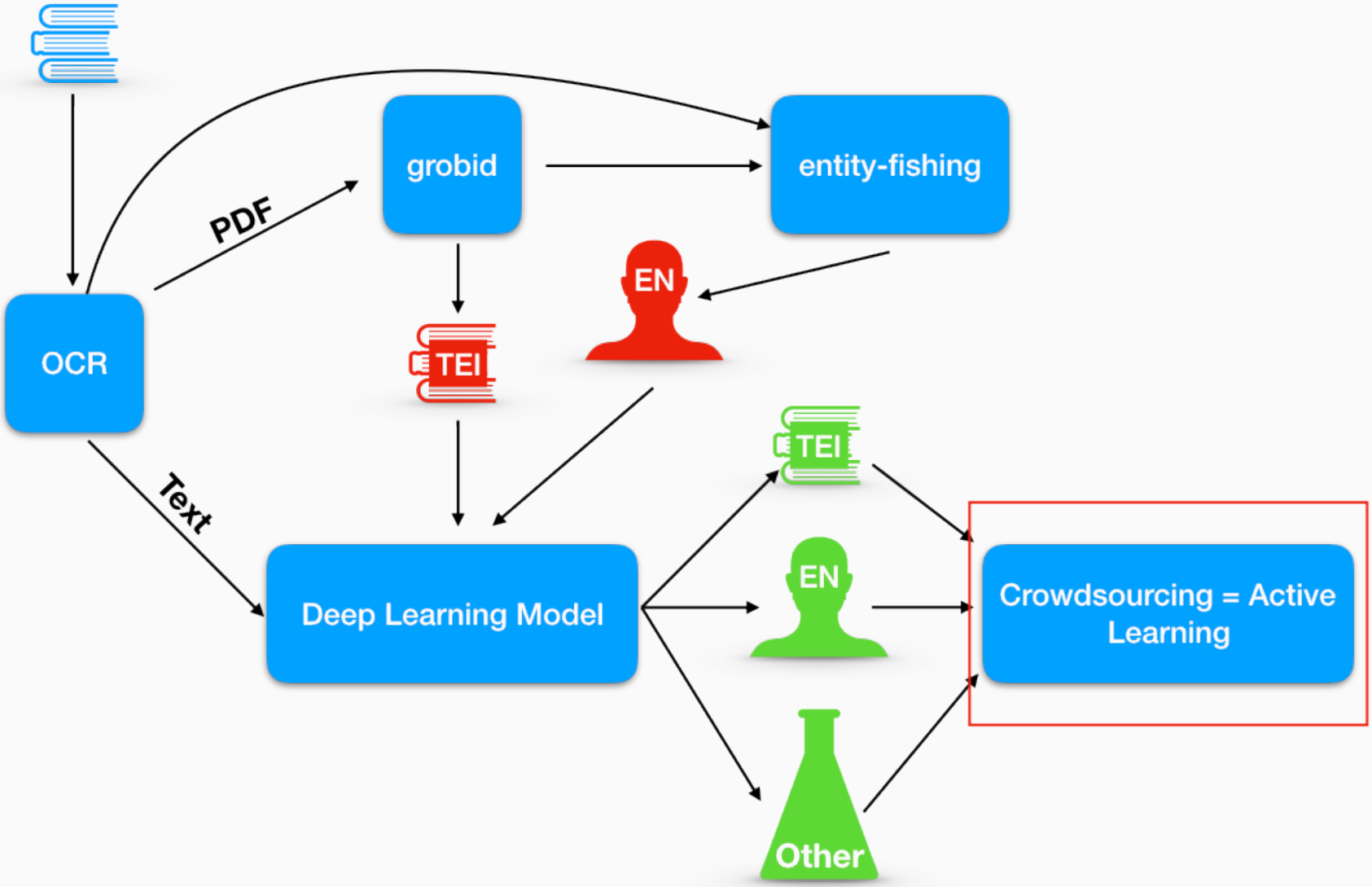
Workflow


grobid
entity-fishing
PDF
OCR
EN
TEI
Text
Deep Learning Model
TEI
EN
Crowdsourcing = Active Learning
Other

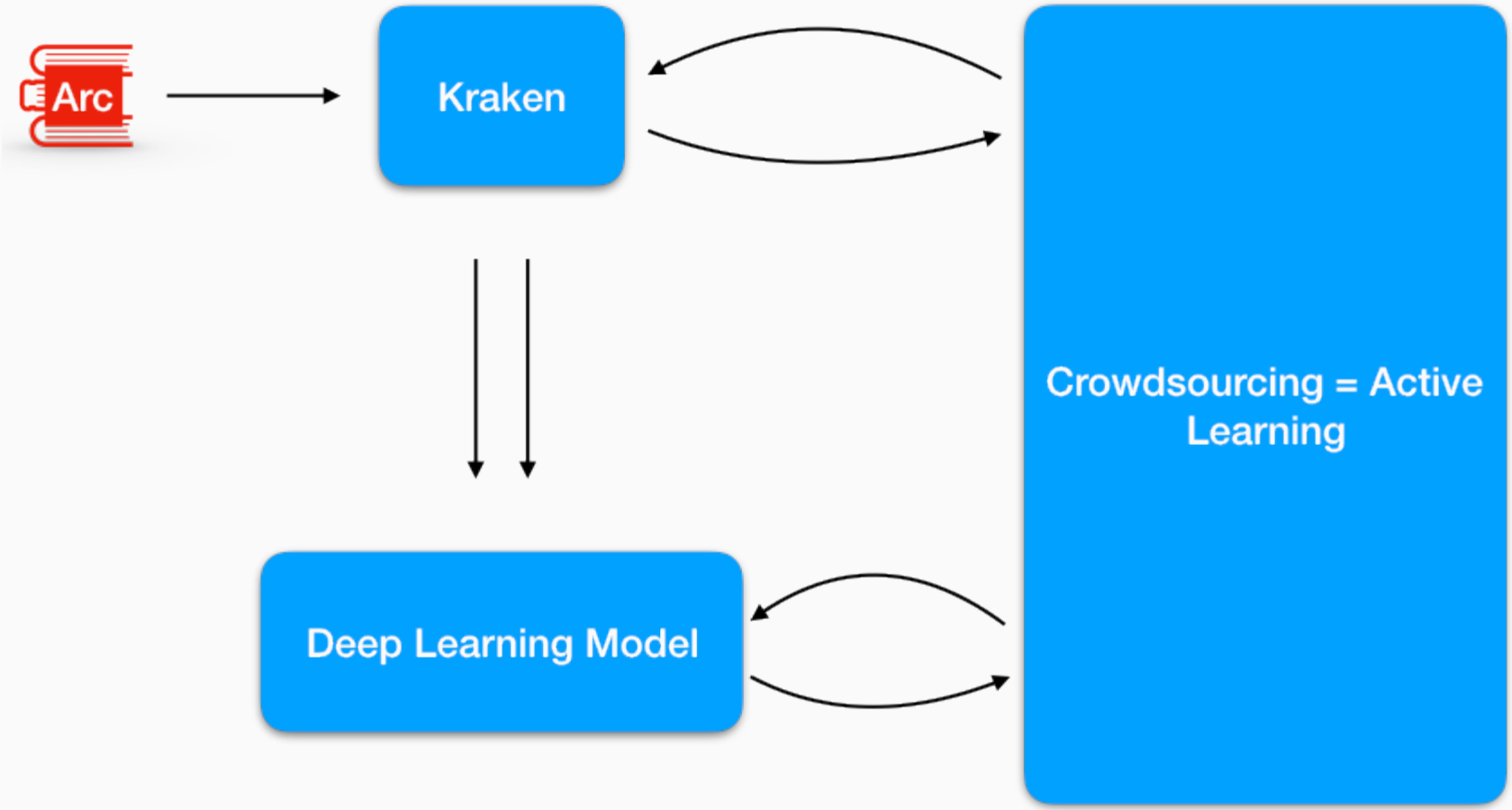
Deep Learning Model



Workflow


grobid
entity-fishing
PDF
EN
OCR
TEI
Text
TEI
Deep Learning Model
EN
Crowdsourcing = Active Learning
Other

Crowdsourcing Model



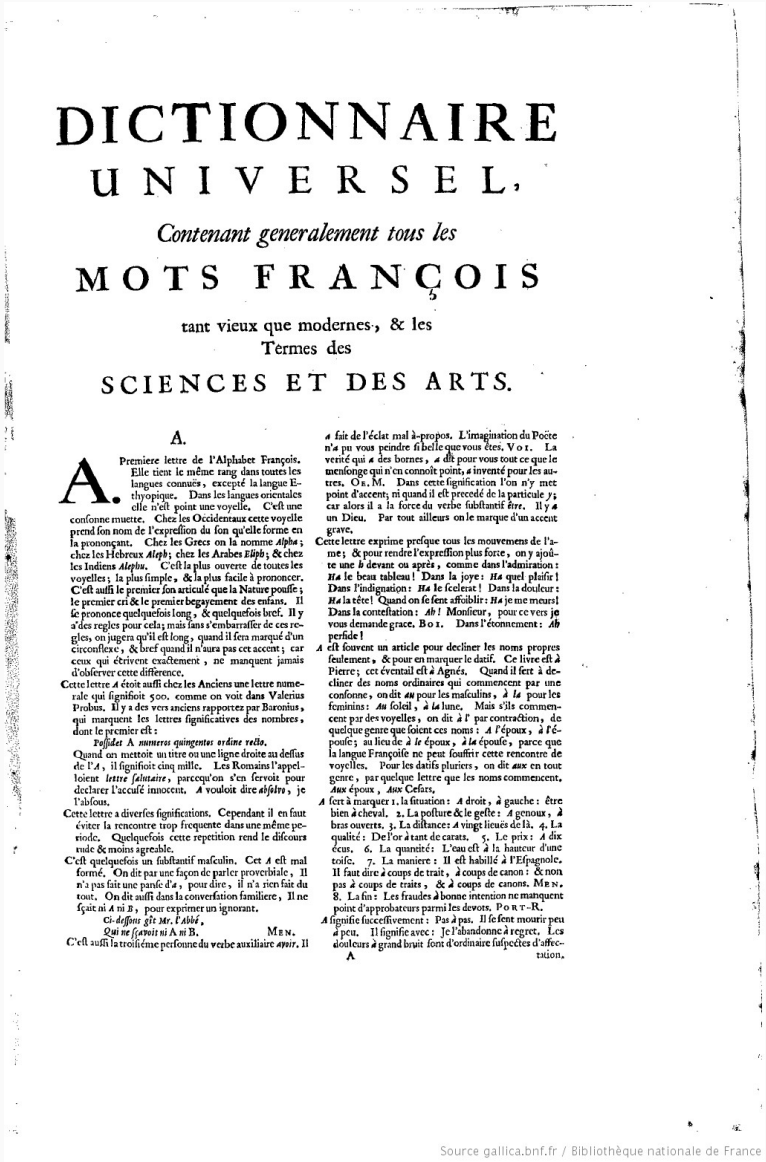
Preparing the Dictionnaire Universel for Automatic Enrichment

Difficulties

- **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!
- **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.

Difficulties

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



DICTIONNAIRE
UNIVERSEL,
Contenant generalement tous les
MOTS FRANÇOIS
tant vieux que modernes, & les
Termes des
SCIENCES ET DES ARTS.

A.

A. Premiere lettre de l'Alphabet François. Elle tient le même rang dans toutes les langues connuës, excepté la langue Ethyopique. Dans les langues orientales elle n'eſt point une voyelle. C'eſt une conſonne muette. Chez les Occidentaux cette voyelle prend ſon nom de l'expreſſion du ſon qu'elle forme en la prononçant. Chez les Grecs on la nomme *Alpha*; chez les Hebreux *Aleph*; chez les Arabes *Eliph*; & chez les Indiens *Alephu*. C'eſt la plus ouverte de toutes les voyelles; la plus ſimple, & la plus facile à prononcer. C'eſt auſſi le premier ſon articulé que la Nature pouſſe; le premier cri & le premier begayement des enfans. Il ſe prononce quelquefois long, & quelquefois bref. Il y a des regles pour cela; mais ſans s'embarraſſer de ces regles, on jugera qu'il eſt long, quand il ſera marqué d'un circonflexe, & bref quand il n'aura pas cet accent; car ceux qui écrivent exactement, ne manquent jamais d'obſerver cette difference.

Cette lettre *A* étoit auſſi chez les Anciens une lettre numerale qui ſignifioit 500. comme on voit dans Valerius Probus. Il y a des vers anciens rapportez par Baronius, qui marquent les lettres ſignificatives des nombres, dont le premier eſt:

Poſſidet A numeros quingentos ordine recto.

Quand on mettoit un titre ou une ligne droite au deſſus de l'*A*, il ſignifioit cinq mille. Les Romains l'appelloient *lettre ſalutaire*, parcequ'on s'en ſervoit pour declarer l'accuſé innocent. *A* vouloit dire *abſolvo*, je l'abſous.

Cette lettre a diverſes ſignifications. Cependant il en faut éviter la rencontre trop frequente dans une même periode. Quelquefois cette repetition rend le diſcours rude & moins agreable.

C'eſt quelquefois un ſubſtantif maſculin. Cet *A* eſt mal formé. On dit par une façon de parler proverbiale, Il n'a pas fait une panſe d'*a*, pour dire, il n'a rien fait du tout. On dit auſſi dans la converſation familiere, Il ne ſçait ni *A* ni *B*, pour exprimer un ignorant.

Ci-deſſous gît Mr. l'Abbé,
Qui ne ſçavoit ni A ni B. MEN.

C'eſt auſſi la troiſiéme perſonne du verbe auxiliaire *avoir*. Il *a* fait de l'éclat mal à-propos. L'imagination du Poëte n'*a* pu vous peindre ſi belle que vous êtes. VOI. La verité qui *a* des bornes, *a* dit pour vous tout ce que le menſonge qui n'en connoît point, *a* inventé pour les autres. OR. M. Dans cette ſignification l'on n'y met point d'accent; ni quand il eſt precedé de la particule *y*; car alors il a la force du verbe ſubſtantif *être*. Il y *a* un Dieu. Par tout ailleurs on le marque d'un accent grave.

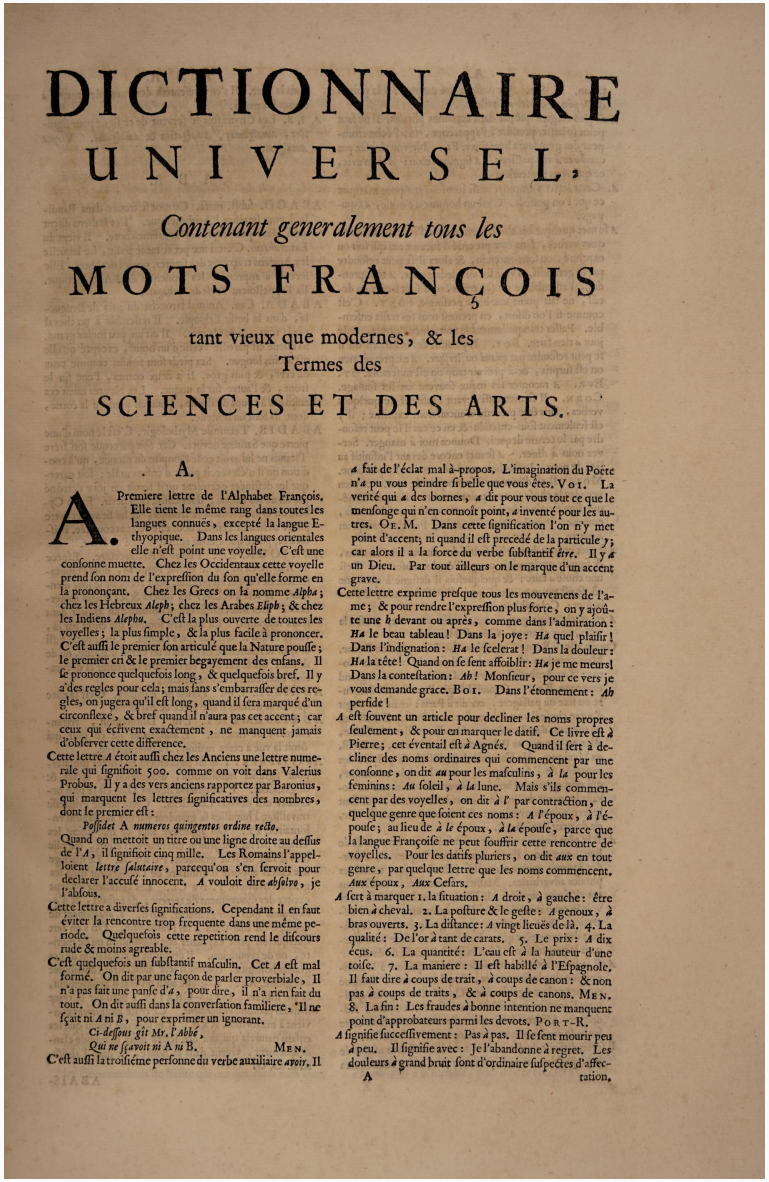
Cette lettre exprime preſque tous les mouvemens de l'ame; & pour rendre l'expreſſion plus forte, on y ajoûte une *h* devant ou après, comme dans l'admiration: *Ha* le beau tableau! Dans la joye: *Ha* quel plaiſir! Dans l'indignation: *Ha* le ſcelerat! Dans la douleur: *Ha* la tête! Quand on ſe ſent affoiblir: *Ha* je me meurs! Dans la conteſtation: *Ah!* Monſieur, pour ce vers je vous demande grace. BOI. Dans l'étonnement: *Ah* perfide!

A eſt ſouvent un article pour decliner les noms propres ſeulement, & pour en marquer le datif. Ce livre eſt *à* Pierre; cet éventail eſt *à* Agnés. Quand il ſert à decliner des noms ordinaires qui commencent par une conſonne, on dit *au* pour les maſculins, *à la* pour les feminins: *Au* ſoleil, *à la* lune. Mais s'ils commencent par des voyelles, on dit *à l'* par contraction, de quelque genre que ſoient ces noms: *A l'*époux, *à l'*épouſe; au lieu de *à le* époux, *à la* épouſe, parce que la langue Françoiſe ne peut ſouffrir cette rencontre de voyelles. Pour les datifs pluriers, on dit *aux* en tout genre, par quelque lettre que les noms commencent. *Aux* époux, *Aux* Ceſars.

A ſert à marquer 1. la ſituation: *A* droit, *à* gauche: être bien *à* cheval. 2. La poſture & le geſte: *A* genoux, *à* bras ouverts. 3. La diſtance: *A* vingt lieuës de là. 4. La qualité: De l'or *à* tant de carats. 5. Le prix: *A* dix écus. 6. La quantité: L'eau eſt *à* la hauteur d'une toiſe. 7. La maniere: Il eſt habillé *à* l'Eſpagnole. Il faut dire *à* coups de trait, *à* coups de canon: & non pas *à* coups de traits, & *à* coups de canons. MEN. 8. La fin: Les fraudes *à* bonne intention ne manquent point d'approbateurs parmi les devots. PORT-R.

A ſignifie ſucceſſivement: Pas *à* pas. Il ſe ſent mourir peu *à* peu. Il ſignifie avec: Je l'abandonne *à* regret. Les douleurs *à* grand bruit ſont d'ordinaire ſuſpectes d'affectation.

A



DICTIONNAIRE
UNIVERSEL,
Contenant generalement tous les
MOTS FRANÇOIS
tant vieux que modernes, & les
Termes des
SCIENCES ET DES ARTS.

A.

A. Premiere lettre de l'Alphabet François. Elle tient le même rang dans toutes les langues connuës, excepté la langue Ethyopique. Dans les langues orientales elle n'eſt point une voyelle. C'eſt une conſonne muette. Chez les Occidentaux cette voyelle prend ſon nom de l'expreſſion du ſon qu'elle forme en la prononçant. Chez les Grecs on la nomme *Alpha*; chez les Hebreux *Aleph*; chez les Arabes *Eliph*; & chez les Indiens *Alephu*. C'eſt la plus ouverte de toutes les voyelles; la plus ſimple, & la plus facile à prononcer. C'eſt auſſi le premier ſon articulé que la Nature pouſſe; le premier cri & le premier begayement des enfans. Il ſe prononce quelquefois long, & quelquefois bref. Il y a des regles pour cela; mais ſans s'embarraſſer de ces regles, on jugera qu'il eſt long, quand il ſera marqué d'un circonflexe, & bref quand il n'aura pas cet accent; car ceux qui écrivent exactement, ne manquent jamais d'obſerver cette difference.

Cette lettre *A* étoit auſſi chez les Anciens une lettre numerale qui ſignifioit 500. comme on voit dans Valerius Probus. Il y a des vers anciens rapportez par Baronius, qui marquent les lettres ſignificatives des nombres, dont le premier eſt:

Poſſidet A numeros quingentos ordine recto.

Quand on mettoit un titre ou une ligne droite au deſſus de l'*A*, il ſignifioit cinq mille. Les Romains l'appelloient *lettre ſalutaire*, parcequ'on s'en ſervoit pour declarer l'accuſé innocent. *A* vouloit dire *abſolvo*, je l'abſous.

Cette lettre a diverſes ſignifications. Cependant il en faut éviter la rencontre trop frequente dans une même periode. Quelquefois cette repetition rend le diſcours rude & moins agreable.

C'eſt quelquefois un ſubſtantif maſculin. Cet *A* eſt mal formé. On dit par une façon de parler proverbiale, Il n'a pas fait une panſe d'*a*, pour dire, il n'a rien fait du tout. On dit auſſi dans la converſation familiere, Il ne ſçait ni *A* ni *B*, pour exprimer un ignorant.

Ci-deſſous gît Mr. l'Abbé,
Qui ne ſçavoit ni A ni B. MEN.

C'eſt auſſi la troiſiéme perſonne du verbe auxiliaire *avoir*. Il *a* fait de l'éclat mal à-propos. L'imagination du Poëte n'*a* pu vous peindre ſi belle que vous êtes. VOI. La verité qui *a* des bornes, *a* dit pour vous tout ce que le menſonge qui n'en connoît point, *a* inventé pour les autres. OR. M. Dans cette ſignification l'on n'y met point d'accent; ni quand il eſt precedé de la particule *y*; car alors il a la force du verbe ſubſtantif *être*. Il y *a* un Dieu. Par tout ailleurs on le marque d'un accent grave.

Cette lettre exprime preſque tous les mouvemens de l'ame; & pour rendre l'expreſſion plus forte, on y ajoûte une *h* devant ou après, comme dans l'admiration: *Ha* le beau tableau! Dans la joye: *Ha* quel plaiſir! Dans l'indignation: *Ha* le ſcelerat! Dans la douleur: *Ha* la tête! Quand on ſe ſent affoiblir: *Ha* je me meurs! Dans la conteſtation: *Ah!* Monſieur, pour ce vers je vous demande grace. BOI. Dans l'étonnement: *Ah* perfide!

A eſt ſouvent un article pour decliner les noms propres ſeulement, & pour en marquer le datif. Ce livre eſt *à* Pierre; cet éventail eſt *à* Agnés. Quand il ſert à decliner des noms ordinaires qui commencent par une conſonne, on dit *au* pour les maſculins, *à la* pour les feminins: *Au* ſoleil, *à la* lune. Mais s'ils commencent par des voyelles, on dit *à l'* par contraction, de quelque genre que ſoient ces noms: *A l'*époux, *à l'*épouſe; au lieu de *à le* époux, *à la* épouſe, parce que la langue Françoiſe ne peut ſouffrir cette rencontre de voyelles. Pour les datifs pluriers, on dit *aux* en tout genre, par quelque lettre que les noms commencent. *Aux* époux, *Aux* Ceſars.

A ſert à marquer 1. la ſituation: *A* droit, *à* gauche: être bien *à* cheval. 2. La poſture & le geſte: *A* genoux, *à* bras ouverts. 3. La diſtance: *A* vingt lieuës de là. 4. La qualité: De l'or *à* tant de carats. 5. Le prix: *A* dix écus. 6. La quantité: L'eau eſt *à* la hauteur d'une toiſe. 7. La maniere: Il eſt habillé *à* l'Eſpagnole. Il faut dire *à* coups de trait, *à* coups de canon: & non pas *à* coups de traits, & *à* coups de canons. MEN. 8. La fin: Les fraudes *à* bonne intention ne manquent point d'approbateurs parmi les devots. PORT-R.

A ſignifie ſucceſſivement: Pas *à* pas. Il ſe ſent mourir peu *à* peu. Il ſignifie avec: Je l'abandonne *à* regret. Les douleurs *à* grand bruit ſont d'ordinaire ſuſpectes d'affectation.

A

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.