MICLE
Micro-Cues of Language Evolution

En dialogue avec les outils d'apprentissage automatique : une chaîne de traitement pour l'annotation syntaxique

Rayan Ziane, ingénieur d'études, CRISCO, Unicaen (projet High-Tech)
Natasha Romanova, coordinatrice de projet, CRISCO, Caen (projet MICLE)
Manon Lavergne, stagiaire M2, CRISCO, Caen (projets High-Tech et MICLE)

UNIVERSITÉ CAEN NORMANDIE

RÉGION NORMANDIE

Centre de Recherches Inter-langues sur la Signification en COntexte
E.A. 4255

AGENCE NATIONALE DE LA RECHERCHE
ANR
DFG

Équipe(s) MICLE et High-Tech

Francfort (Vénitien)
Cecilia Poletto (PI)
Francesco Pinzin

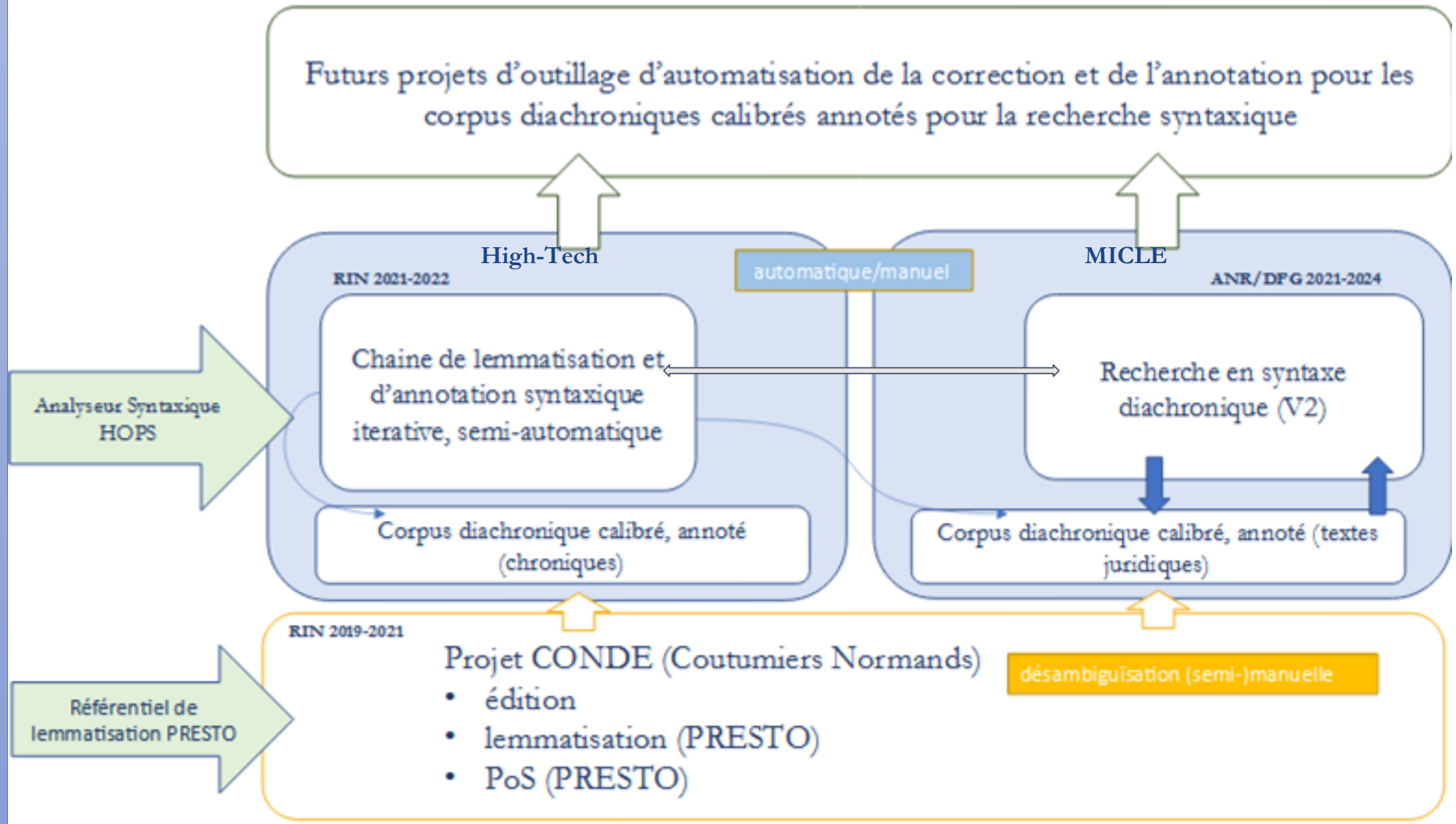
Stagiaires:
Leah Pavcic
Francesca Santangelo

Caen (Français de Normandie)
Pierre Larrivée (PI)
Mathieu Goux
Natasha Romanova
Rayan Ziane

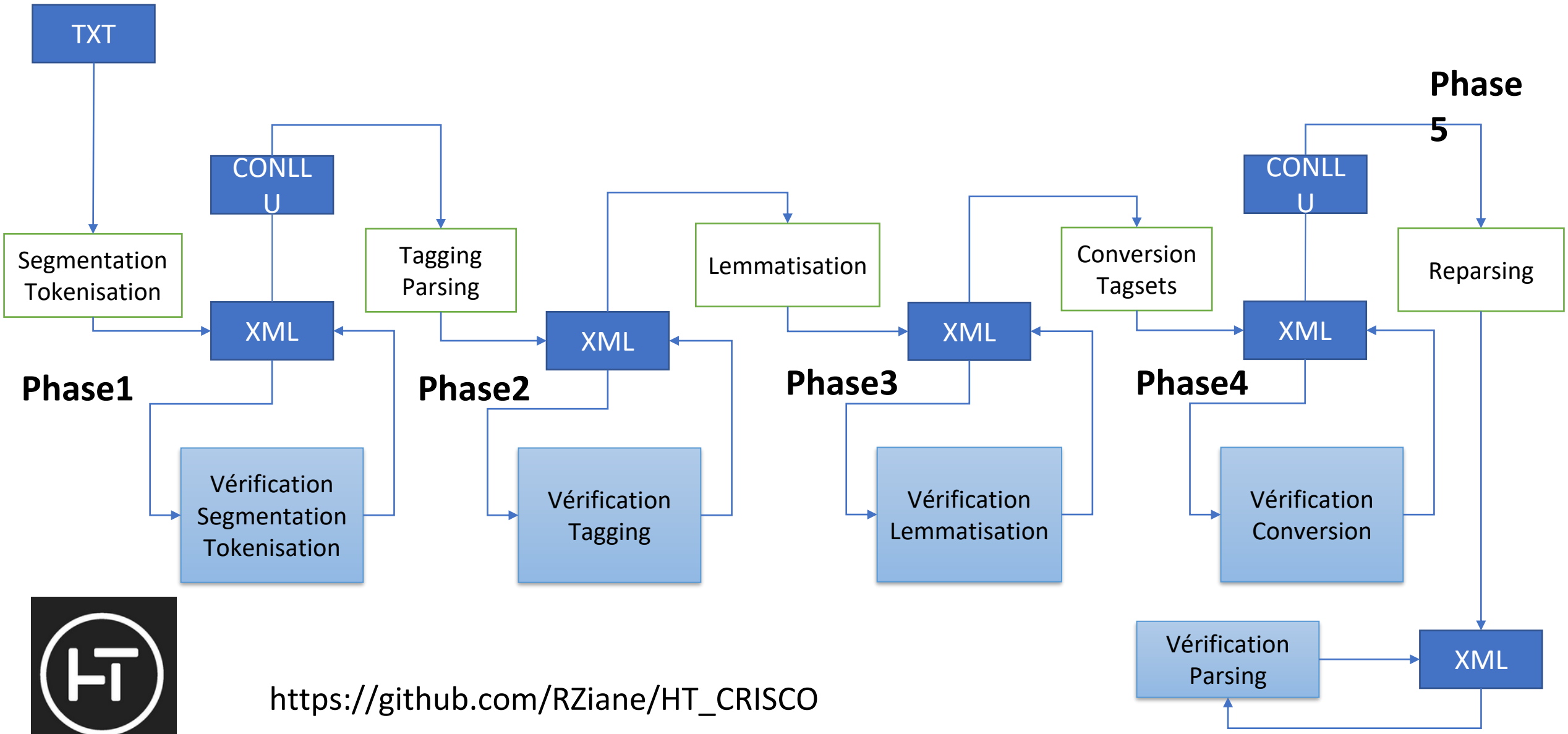
Stagiaires:
Agathe Aubert
Manon Lavergne
Lucy Marie-Leblanc
Marie Picart
Valentin Simenel
Yichu Wang

MICLE	06/2021 – 05/2024 | DFG/ANR
High-Tech	11/2021 - 10/2023 | RIN (Réseau d'intérêts normands)

MICLE
Micro-Cues of Language Evolution



Workflow



https://github.com/RZiane/HT_CRISCO

1/ Segmentation et tokenisation

Fichier texte

Deux formes possibles:

- Structure (pages et paragraphes)
- Structure + information typographique

TXT

Script python

Tokenisation:

- Par espace blanc
- Par mot lexical

Segmentation « stratégique » (en phrase):

- Par point dans l'édition

Tokenisation
Segmentation

XML

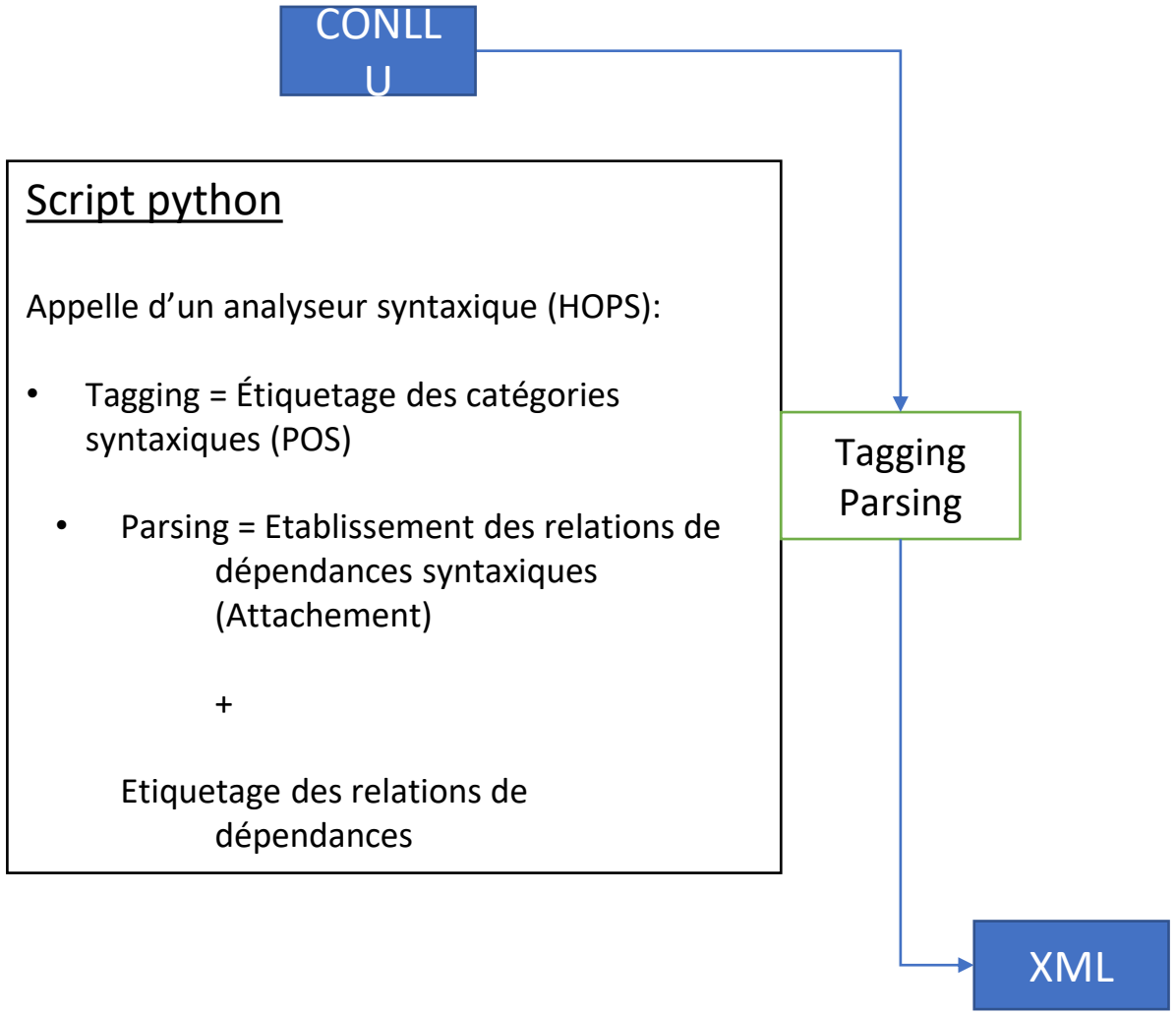
Le format texte (.txt):

```
L'histoire de la Normandie paraît ne commencer
qu'à l'époque où, devenue par la conquête des
hommes du Nord ou Scandinaves, une souverai¬
neté particulière, elle offre à la postérité de grands
traits et des événements dignes de remarque.
Cependant une esquisse rapide des siècles qui
ont précédé l'invasion des pirates de la Norvège
nous a paru devoir former la matière du premier
```

Le format eXtensible Markup Language (.xml):

```
<s n="1">
    <w n="1">L'</w>
    <w n="2">histoire</w>
    <w n="3">de</w>
    <w n="4">la</w>
    <w n="5">Normandie</w>
    <w n="6">parait</w>
    <w n="7">ne</w>
    <w n="8">commencer</w>
    <w n="9">qu'</w>
    <w n="10">à</w>
    <w n="11">l'</w>
    <w n="12">époque</w>
    <w n="13">où</w>
    <w n="14">,</w>
    <w n="15">devenue</w>
```

2/ Tagging et Parsing



- HOPS (Grobol et Crabbé, 2021)
- UD (de Marneffe et al., 2021)

Les formats de fichier (2):

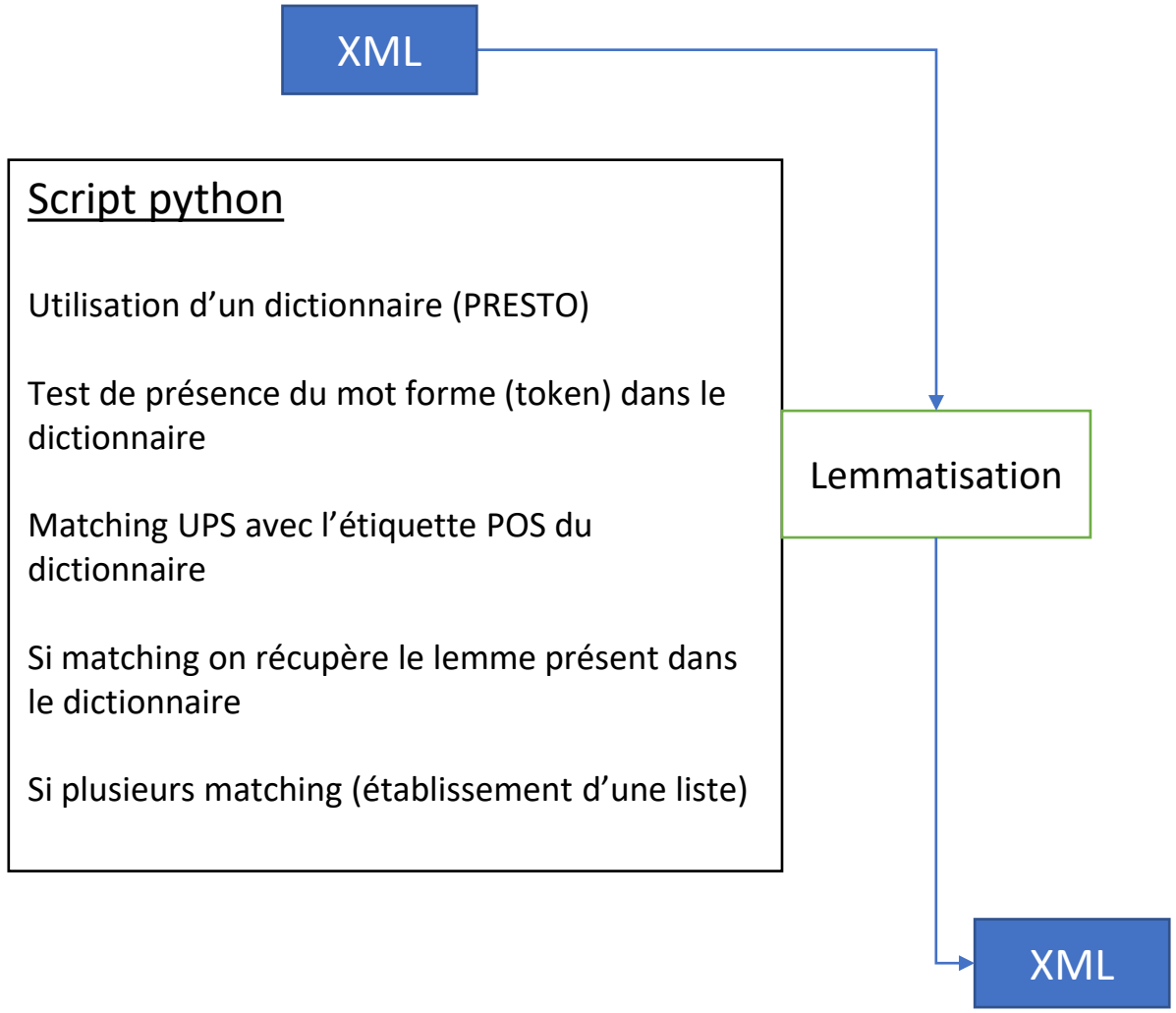
Le format conll (.txt):

```
# sent_id = 1-1-1-1-1
1   L'          _   DET     _   _   2   det     _   join=_
2   histoire    _   NOUN    _   _   6   nsubj   _   join=_
3   de          _   ADP     _   _   5   case    _   join=_
4   la          _   DET     _   _   5   det     _   join=_
5   Normandie   _   PROPN   _   _   2   nmod    _   join=_
6   parait      _   VERB    _   _   0   root    _   join=_
7   ne          _   ADV     _   _   8   advmod  _   join=_
8   commencer   _   VERB    _   _   6   xcomp   _   join=_
9   qu'         _   ADV     _   _   8   mark    _   join=_
10  à           _   ADP     _   _   12  case    _   join=_
```

Le format eXtensible Markup Language (.xml):

```
<s n="1">
  <w n="1" udpos="DET" lemma="_" head="2" function="det">L'</w>
  <w n="2" udpos="NOUN" lemma="_" head="6" function="nsubj">histoire</w>
  <w n="3" udpos="ADP" lemma="_" head="5" function="case">de</w>
  <w n="4" udpos="DET" lemma="_" head="5" function="det">la</w>
  <w n="5" udpos="PROPN" lemma="_" head="2" function="nmod">Normandie</w>
  <w n="6" udpos="VERB" lemma="_" head="0" function="root">paraît</w>
  <w n="7" udpos="ADV" lemma="_" head="8" function="advmod">ne</w>
  <w n="8" udpos="VERB" lemma="_" head="6" function="xcomp">commencer</w>
  <w n="9" udpos="ADV" lemma="_" head="12" function="mark">qu'</w>
  <w n="10" udpos="ADP" lemma="_" head="12" function="case">à</w>
```

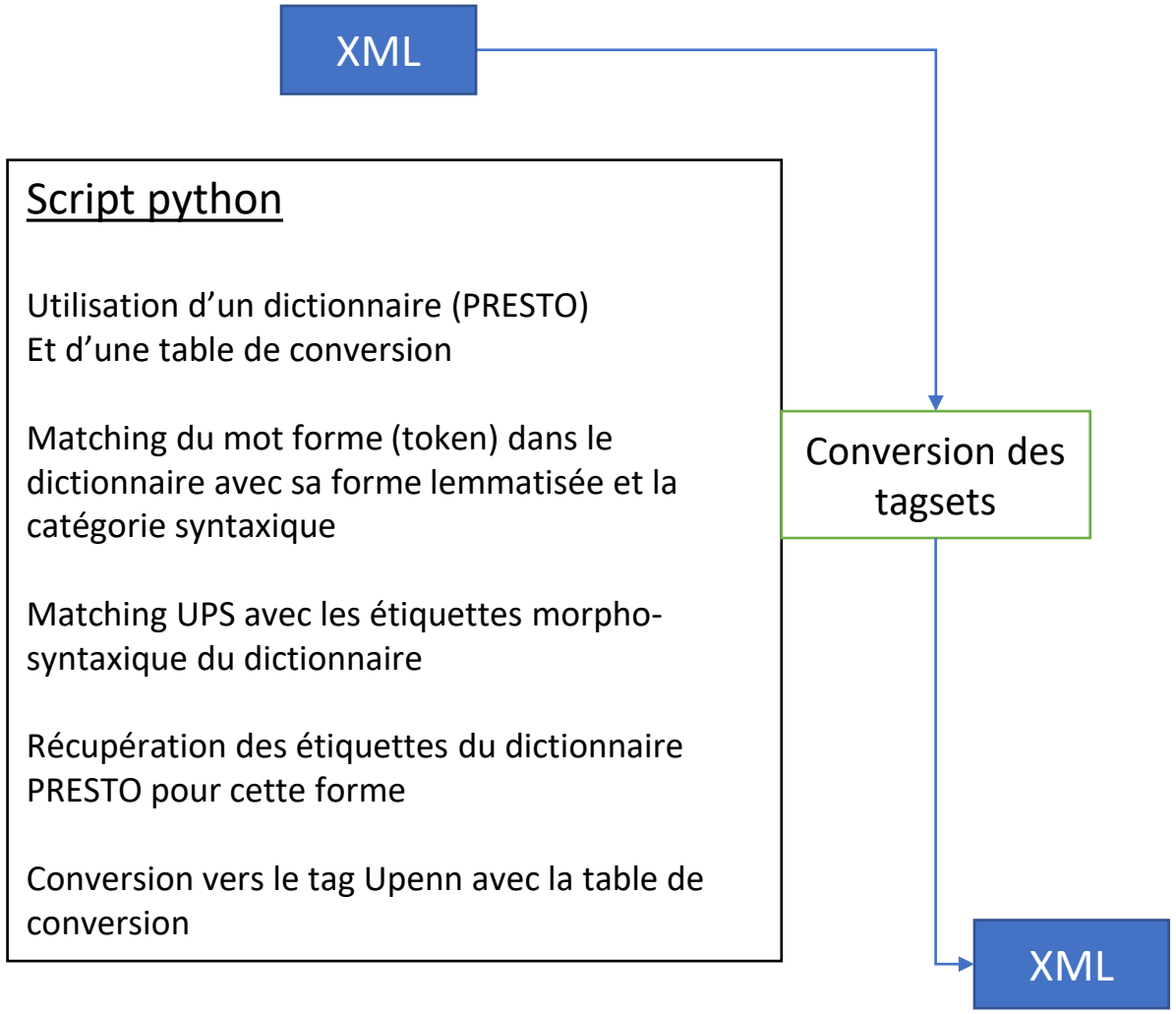
3/ Lemmatisation



- PRESTO (Blumenthal et al, 2017)

```
les/PRO/Pp/IL
les/DET/Da/LE
les/PREP/S/LÈS
les/ADJ/Ag/LÉ
les/NOM/Nc/LÉ
```

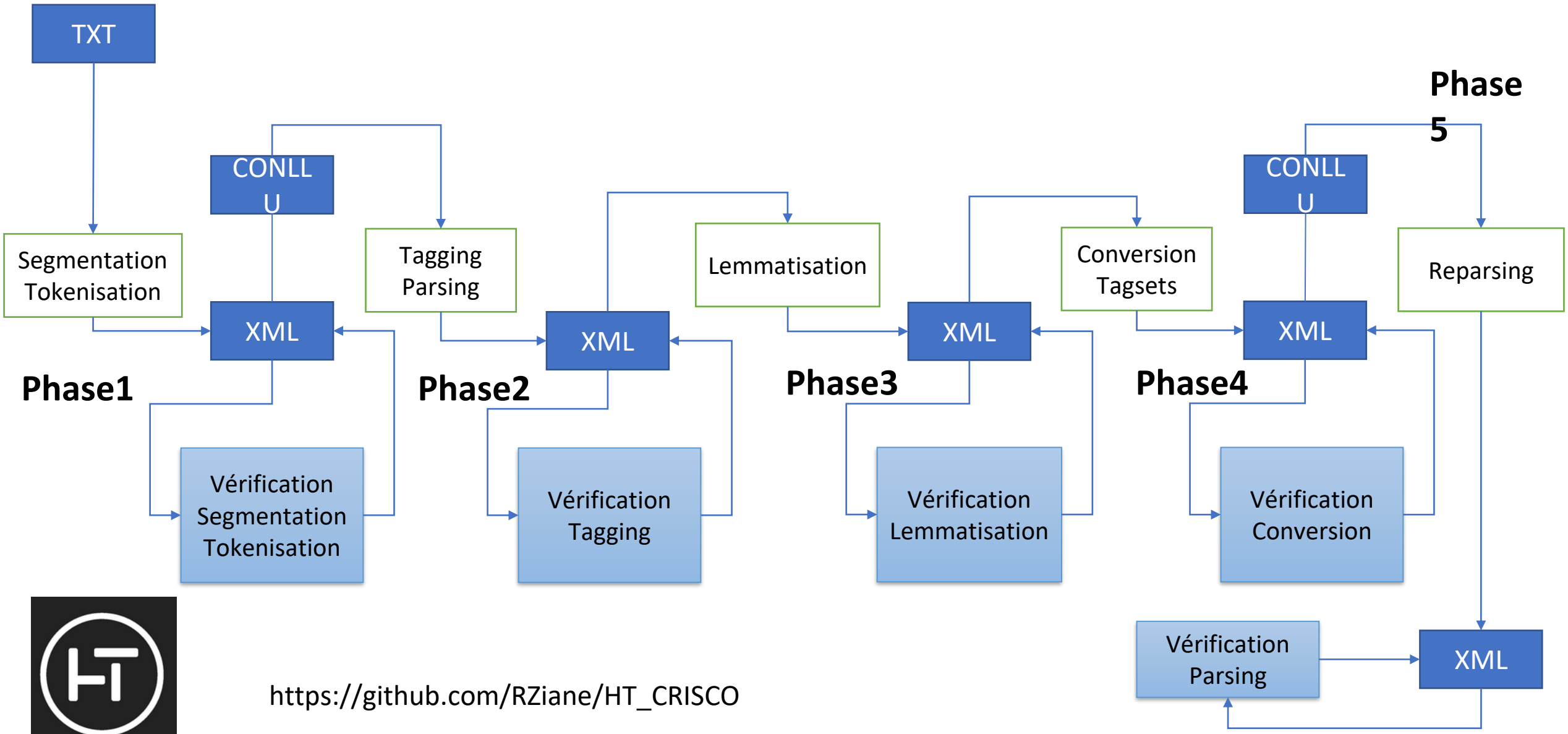
4/ Conversion des tagsets



- UPENN (Santorini, 2007)

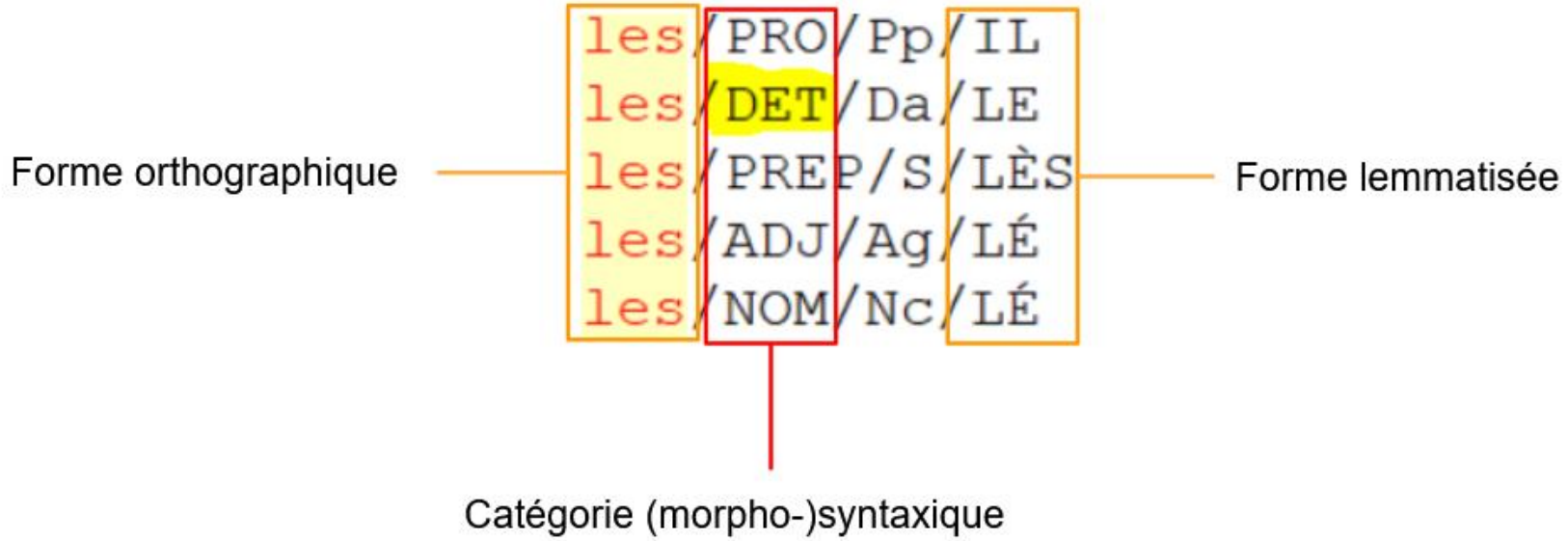
```
fait/PAG/Ge/FAIRE
fait/VER/Vvc/FAIRE
fait/ADJ/Ag/FAIT
fait/NOM/Nc/FAIT
```

Workflow



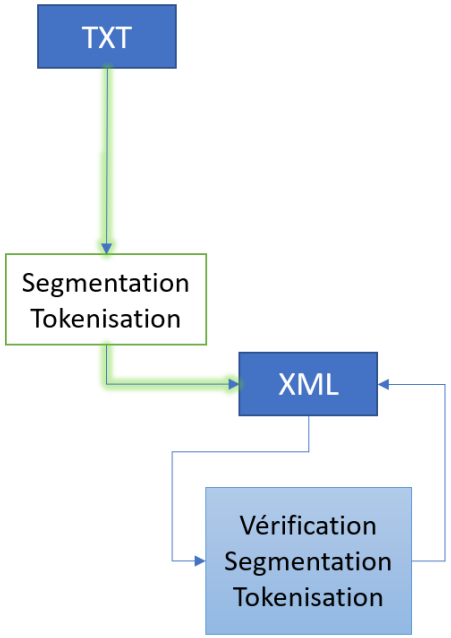
https://github.com/RZiane/HT_CRISCO

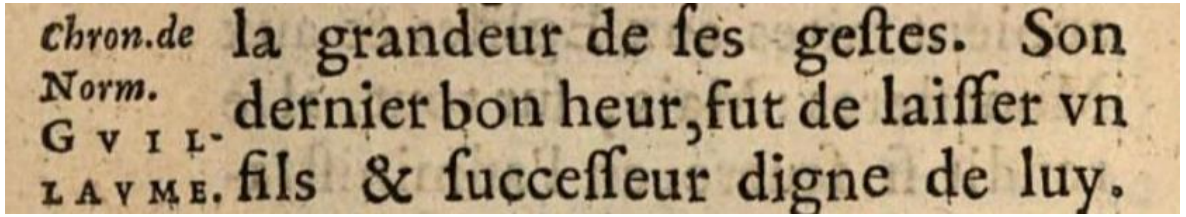
H-T-CRISCO: principe d'enchaînement des informations



Blumenthal, P., Diwersy, S., Falaise, A., Lay, H., Souvay, G., Vigier, D., Descartes, P. R., Nancy, U., & de Lyon, U. (2017). *Presto, un corpus diachronique pour le français des XVIe-XXe siècles*.

H-T-CRISCO Phase 1 (Seg/Tok)




chron.de Norm. GVILLAVME. la grandeur de ſes geſtes. Son dernier bon heur, fut de laiſſer vn fils & ſucceſſeur digne de luy.

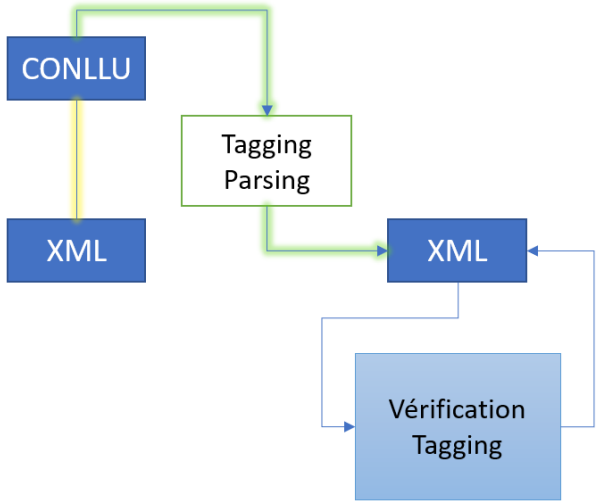
Son dernier bon heur, fut de laisser vn fils

```
<s n="13">
    <w n="1">Son</w>
    <w n="2">dernier</w>
    <w n="3">bon</w>
    <w n="4">heur</w>
    <w n="5">,</w>
    <w n="6">fut</w>
    <w n="7">de</w>
    <w n="8">laisser</w>
    <w n="9">vn</w>
    <w n="10">fils</w>
```

```
<s n="13">
    <w n="1">Son</w>
    <w n="2">dernier</w>
    <w n="3">
        <choice>
            <sic>bon heur,</sic>
            <corr>bonheur</corr>
        </choice>
    </w>
    <w n="6">fut</w>
    <w n="7">de</w>
    <w n="8">laisser</w>
    <w n="9">vn</w>
    <w n="10">fils</w>
```

Jacques-Eustache d'Anneville. 1665. Abrégé de l'histoire de Normandie. Rouen : Jacques et Jean Lucas. (H-T)

H-T-CRISCO Phase 2 (Tagging PoS UD)

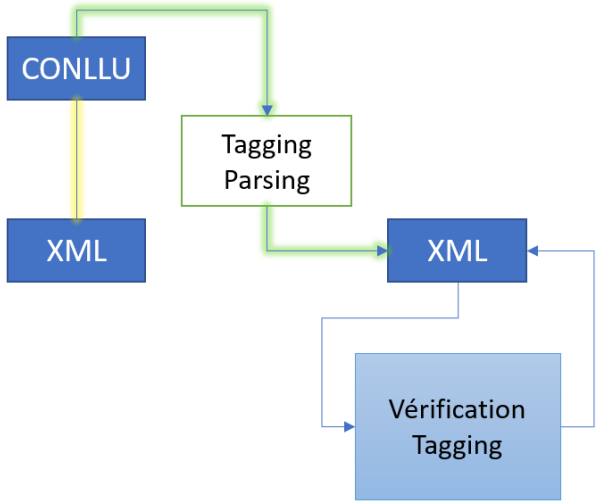



```
<w n="13" udpos="DET" lemma="_" head="14" function="det">le</w>
<w n="14" udpos="NOUN" lemma="_" head="8" function="conj">choix</w>
<w n="15" udpos="ADP" lemma="_" head="16" function="case:det">des</w>
<w n="16" udpos="PROPN" lemma="_" head="14" function="nmod">Nations</w>
```

4661	PROPN	Narbonnoise	
4662	PROPN	Nations	NOUN
4663	PROPN	Nauale	ADJ
4664	PROPN	Nauarre	
4665	PROPN	Nauarrois	
4666	PROPN	Nauires	NOUN
4667	PROPN	Neel	
4668	PROPN	Nemours	
4669	PROPN	Neuers	

Jacques-Eustache d'Anneville. 1665. Abrégé de l'histoire de Normandie. Rouen : Jacques et Jean Lucas. (H-T)

H-T-CRISCO Phase 2 (Tagging PoS UD)



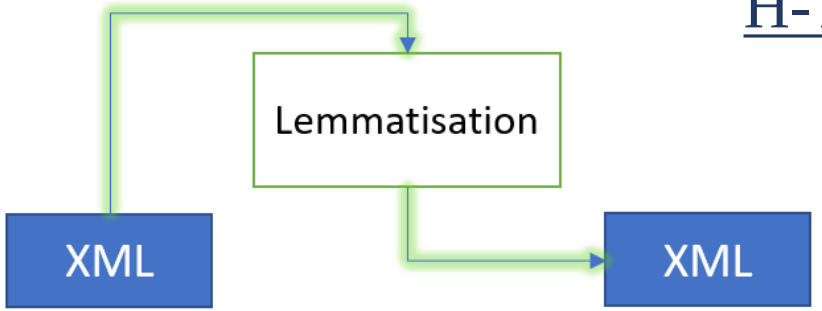
ains s'en ala viers Roumenel

```
<w n="22" head="25" function="advmod" udpos="ADV">ams</w>
<w n="23" head="25" function="expl" udpos="PRON">s'</w>
<w n="24" head="25" function="obl" udpos="ADV">en</w>
<w n="25" head="18" function="conj" udpos="VERB">ala</w>
<w n="26" head="27" function="case" udpos="ADP">viers</w>
<w n="27" head="25" function="obl" udpos="PROPN">Roumenel</w>
```

ADV	amont
ADV	ams
ADV	anciienement

F. Michel. 1840. *Histoire des Ducs de Normandie et des Rois d'Angleterre*, Paris : Renouard. (13ème) (H-T)

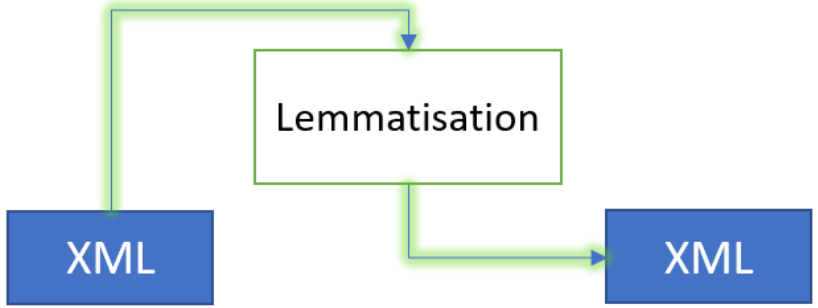
H-T-CRISCO Phase 3 (Lemmatisation)




```
<w function="cc.nc" head="3" n="1" udpos="CCONJ" lemma="et">Et</w>
<w function="advmod" head="3" n="2" udpos="ADV" lemma="lors">lors</w>
<w function="root" head="0" n="3" udpos="VERB" lemma="avoir///oser///ouïr">ot</w>
<w function="det" head="5" n="4" udpos="DET" lemma="le">le</w>
<w function="nsubj" head="3" n="5" udpos="NOUN" lemma="comte///conte///conté">conte</w
<w function="case" head="7" n="6" udpos="ADP" lemma="de">de</w>
<w function="flat" head="5" n="7" udpos="PROPN" lemma="flandres">Flandres</w>
<w function="obj" head="3" n="8" udpos="NOUN" lemma="conseil">conseil</w>
<w function="case" head="10" n="9" udpos="ADP" lemma="de">d'</w>
<w function="nmod" head="8" n="10" udpos="PRON" lemma="aucun">aucuns</w>
<w function="case" head="13" n="11" udpos="ADP" lemma="de">de</w>
<w function="det" head="13" n="12" udpos="DET" lemma="son">ses</w>
<w function="nmod" head="10" n="13" udpos="NOUN" lemma="admis///ami///ammi">amys</w>
```

Auguste et Emile Molinier. 1882. Chronique normande du XIVe siècle. Paris: Renouard. (H-T)

H-T-CRISCO Phase 3 (Lemmatisation)




```
<w function="cc" head="32" n="30" udpos="CCONJ" lemma="et">et</w>
<w function="iobj" head="32" n="31" udpos="PRON" lemma="il">lui</w>
<w function="conj" head="11" n="32" udpos="VERB" lemma="commander">commanda</w>
<w function="mark" head="35" n="33" udpos="SCONJ" lemma="que">que</w>
<w function="nsubj" head="35" n="34" udpos="PRON" lemma="il">il</w>
<w function="ccomp" head="32" n="35" udpos="VERB" NoMatchingPresto="Word" lemma="_">vuydast</w>
<w function="case" head="38" n="36" udpos="ADP" lemma="de">de</w>
<w function="det" head="38" n="37" udpos="DET" lemma="le">la</w>
<w function="obl" head="35" n="38" udpos="NOUN" lemma="conté">conté</w>
<w function="case" head="40" n="39" udpos="ADP" lemma="de">de</w>
<w function="nmod" head="38" n="40" udpos="PROPN" lemma="flandres">Flandres</w>
```

Auguste et Emile Molinier. 1882. Chronique normande du XIVe siècle. Paris: Renouard. (H-T)

H-T-CRISCO Phase 3 (Lemmatisation)

```
<w function="nsubj" head="10" n="8" udpos="PRON" lemma="qui">qui</w>
<w function="aux" head="10" n="9" udpos="AUX" lemma="ester///être">estoit</w>
<w function="acl:relcl" head="4" n="10" udpos="VERB" lemma="_" NoMatchingPresto="POS">nepveu</w>
<w function="case:det" head="12" n="11" udpos="ADP" lemma="à+le">au</w>
<w function="obl" head="10" n="12" udpos="NOUN" lemma="comte///conte///conté">conte</w>
<w function="case" head="14" n="13" udpos="ADP" lemma="de">de</w>
<w function="flat" head="12" n="14" udpos="PROPN" lemma="flandres">Flandres</w>
```

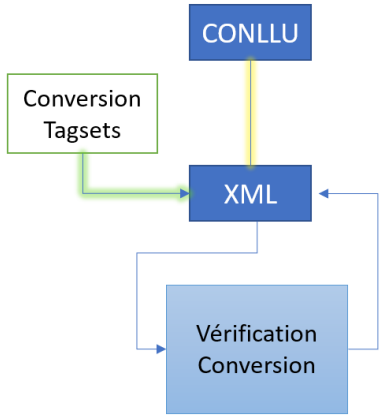
```
nepueu/NOM/Nc/NEVEU
nepveu/NOM/Nc/NEVEU
```

```
<w function="aux:pass" head="14" n="12" udpos="AUX" lemma="ester///être">estoit</w>
<w function="advmod" head="14" n="13" udpos="ADV" lemma="moult">moult</w>
<w function="ccomp" head="3" n="14" udpos="VERB" lemma="_" NoMatchingPresto="POS">convoiteux</w>
<w function="punct" head="29" n="15" udpos="PUNCT" lemma=",">,</w>
```

```
convoiteux/ADJ/Ag/CONVOITEUX
convoiteux/NOM/Nc/CONVOITEUX
```

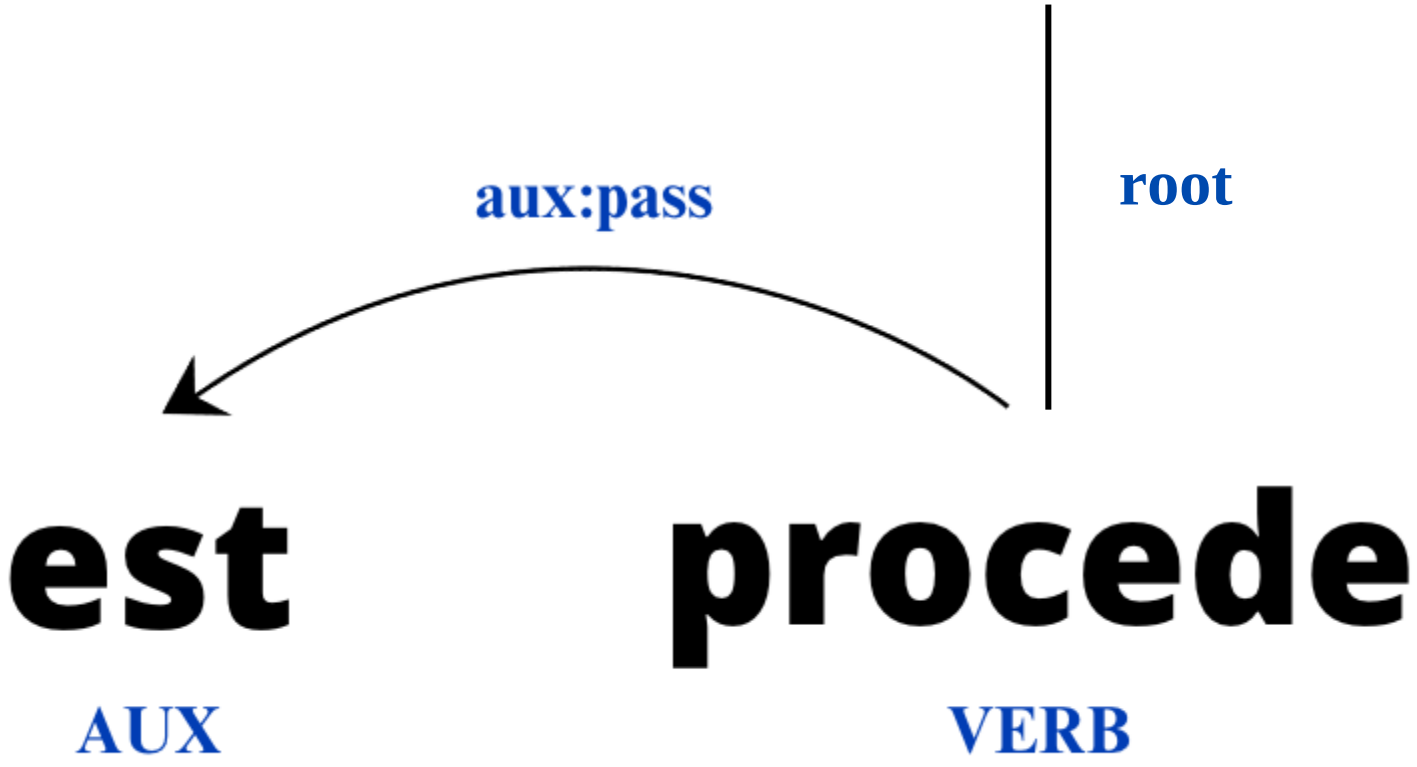
Auguste et Emile Molinier. 1882. Chronique normande du XIVe siècle. Paris: Renouard (H-T)

H-T-CRISCO Phase 4 (conversion PoS UPenn/PRESTO)




```
<w join="_" n="14" head="20" function="nsubj" lemma="celui" udpos="PRON" prpos="Pd" uppos="PRO">celuy</w>
<w join="_" n="15" head="17" function="nsubj" lemma="qui" udpos="PRON" prpos="Pr" uppos="WPRO" ambiguite="PronType">qui</w>
<w join="_" n="16" head="17" function="obl" lemma="le" udpos="PRON" prpos="_" uppos="_" NoMatchingPresto="Word">l'</w>
<w join="_" n="17" head="14" function="acl:relcl" lemma="envoyer" udpos="VERB" prpos="Vvc" uppos="VJ" ambiguite="standard">enuoye</w>
```

Guillaume Terrien, Commentaire du droit civil tant public que privé observé en Normandie. 1578 (MICLE)


aux:pass
root
est
procede
AUX
VERB

MICLE
Micro-Cues of Language Evolution

```
<w udpos="CCONJ" head="3" function="cc:nc" lemma="et" n="1" prpos="Cc" uppos="CONJO">Et</w>LF
<w udpos="AUX" head="3" function="aux:pass" lemma="être" n="2" prpos="Vuc" uppos="EJ">est</w>LF
<w udpos="VERB" head="0" function="root" lemma="procéder" n="3" prpos="Ge" uppos="VPP" ambiguite="standard">procede</w>LF
<w udpos="ADP" head="5" function="case" lemma="en" n="4" prpos="S" uppos="P">en</w>LF
<w udpos="NOUN" head="3" function="obl" lemma="bailliage" n="5" prpos="Nc" uppos="NCS">bailliage</w>LF
```

```
<w udpos="AUX" head="5" function="aux:pass" lemma="avoir" n="3" prpos="Vuc" uppos="AJ">a</w>LF
<w udpos="AUX" head="5" function="aux" lemma="être" n="4" prpos="Ge" uppos="EPP" ambiguite="auxiliaire">este</w>LF
<w udpos="VERB" head="0" function="root" lemma="faire" n="5" prpos="Ge" uppos="VPP" ambiguite="standard">faict</w>LF
```

```
<w udpos="DET" head="10" function="det" lemma="le" n="9" prpos="Da" uppos="D">les</w>LF
<w udpos="NOUN" head="5" function="obl" lemma="ordonnance" n="10" prpos="Nc" uppos="NCS">ordonnances</w>LF
```

Bibliographie

Blumenthal, P., Diwersy, S., Falaise, A., Lay, H., Souvay, G., Vigier, D., Descartes, P. R., Nancy, U., & de Lyon, U. (2017). Presto, un corpus diachronique pour le français des XVIe-XXe siècles.

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. Computational Linguistics, 47(2), 255-308. https://doi.org/10.1162/coli_a_00402

DeRose, S. (1999). XML and the TEI. Computers and the Humanities, 33(1), 11-30. https://doi.org/10.1023/A:1001771114509

Goux, M. & Pinzin F. Challenges of a Multilingual Corpus (Old French/Old Venetian): The Example of the MICLE project. Venise et la France. Similitudes, spécificités, interrelations. Castro E., Della Fontana A. and Pezzini E. Franco Cesati (ed) Florence : Cesati Editore (sous presse).

Grobol, L., & Crabbé, B. (2021). Analyse en dépendances du français avec des plongements contextualisés (French dependency parsing with contextualized embeddings). Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale, 106-114. https://aclanthology.org/2021.jeptalnrecital-taln.9

Lay, M.-H. & Pincemin, B. Pour une exploration humaniste des textes : AnaLog. Statistical Analysis of Textual Data: Proceedings of 10th International Conference Journée d'Analyse statistique des Données Textuelles 9-11 Juin 2010 – Sapienza University of Rome. Bolasco, S., Chiari I. & Giuliano L. (eds) V.2, 1045-1056 https://www.ledonline.it/ledonline/JADT-2010/allegati/JADT-2010-1045-1056_106-Lay.pdf

Morcos, H., Noël, G. & Husar, M. Lemmatization in the collaborative editorial workflow of a medieval French text: The digital edition of the Ancient History jusqu'à César. Digital Scholarship in the Humanities, 36(2), 203-209. https://doi.org/10.1093/llc/fqaa060

Santorini, B. (2007) Protocole d'étiquetage - Parties du discours (PDD). https://www.ling.upenn.edu/~beatrice/corpus-ling/annotation-french/pos/pos-index.html