Comment construire un corpus pour l’analyse en linguistique historique

Helena Bermúdez Sabel

helena.bermudez@unine.ch

Linguistique de corpus: récapitulation

Méthodologie (Wallis & Nelson 2001):

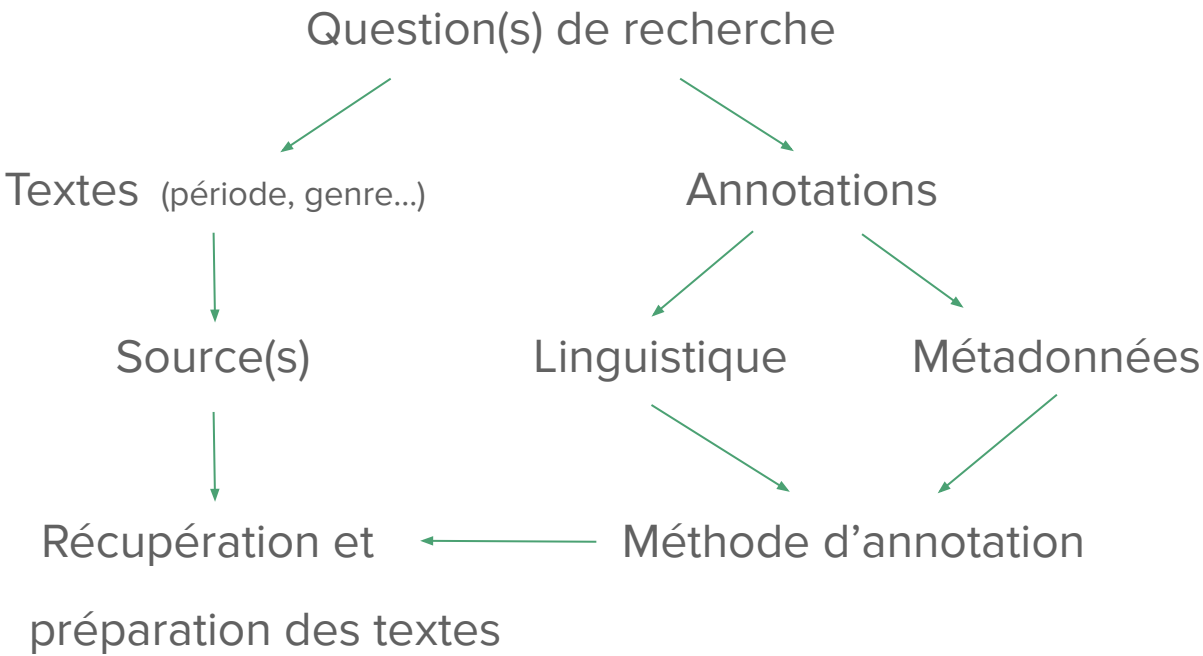
- **Annotation**: application d'un schéma à une série de textes (*cf*. Sinclair 2004)
- **Abstraction**: mise en correspondance entre les éléments d'un schéma et les éléments d'un modèle théorique ou d'un ensemble de données. Apprentissage de règles, formulation d'hypothèses.
- **Analyse**: exploration statistique, inférences, généralisations à partir du jeu de données.

Avant de commencer

- **Modélisation égoïste**: pour exprimer des idées de recherche spécifiques dans les cas où des données sont créées pour répondre aux propres besoins de recherche du créateur
- **Modélisation altruiste**: pour servir de format d'échange pour certains types d'utilisateurs et communautés d'utilisateurs où les données sont généralement créées et modélisées en tenant compte des besoins de quelqu'un d'autre

(Jannidis & Flanders 2013)

Avant de commencer



Avant de commencer

Quelles ressources ai-je? (quels outils puis-je utiliser?

quelles sont mes compétences (en programmation)? combien de temps ai-je?)

Récupération et préparation des textes

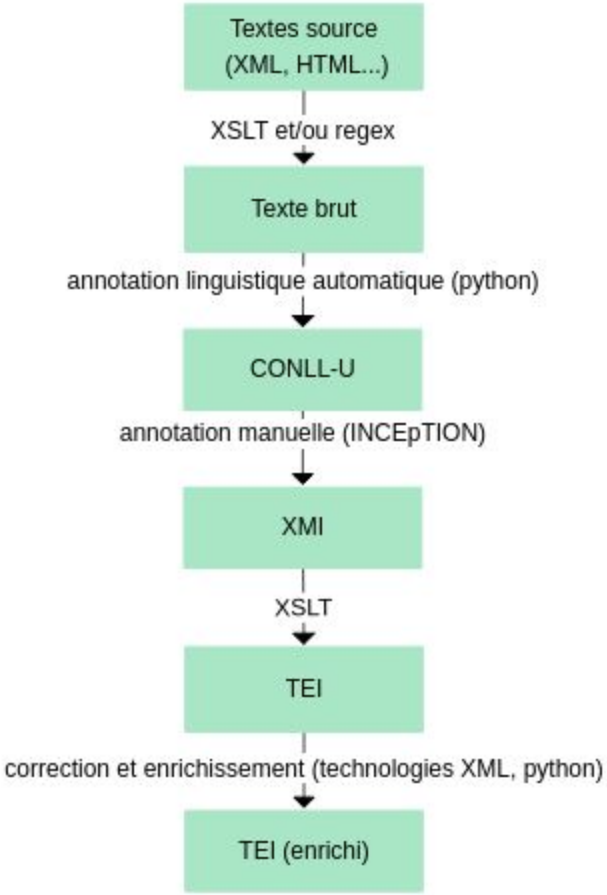
Méthode d’annotation

WoPoss: aperçu préliminaire

- **Question de recherche**: quelle est l'évolution des marqueurs modaux en latin sur une longue période (1000 ans)?
- **Textes**: *cf. critères de représentativité expliqués par Prof. Dell'Oro*
- **Schéma d'annotation**: lemme, partie du discours, caractéristiques morphologiques, dépendances syntaxiques, modalité (marqueur modal, portée, typologie, participant ...)
- **Méthode d'annotation**: annotation automatique + manuelle
- **Outils**: librairies python TAL, INCEpTION, eXist-DB
- **Ressources**: financement de 5 ans, équipe de 3 membres...

WoPoss: exigences

- Représentativité du corpus >> sources multiples (hétérogènes)
- Outils >> pipeline qui commence par du texte brut et se termine par un XML fortement annoté


Textes source
(XML, HTML...)
XSLT et/ou regex
Texte brut
annotation linguistique automatique (python)
CONLL-U
annotation manuelle (INCEpTION)
XMI
XSLT
TEI
correction et enrichissement (technologies XML, python)
TEI (enrichi)

Préparation du corpus

Après la sélection des œuvrages:

- Récupération de documents à partir de différentes sources en ligne
- Homogénéisation

Récupération de textes

Défis:

- Droits d’auteur
- Qualité philologique de la source
- Fichiers source dans différents formats: HTML, TXT, XML-TEI (Epidoc ou autre personnalisation TEI)

Sources: exemple

```
<ab>
    <lb n="1"/>
    <expan>
        <abbr>D</abbr>
        <ex>is</ex>
    </expan>
    <expan>
        <abbr>M</abbr>
        <ex>anibus</ex>
    </expan>.<lb n="2"/>
    <expan>
        <abbr>L</abbr>
        <ex>ucius</ex>
    </expan> Pullaenu<lb break="no" n="3"/>s Zosimus<lb n="4"/>fecit coìu<lb break="no" n="5"/>
    <choice>
        <sic>gi</sic>
        <corr>coniugi</corr>
    </choice>
    <choice>
        <sic>
            <expan>
                <abbr>kar</abbr>
                <ex>issimae</ex>
```

Sources: exemple

```
<html>
 <head>
  <meta http-equiv="content-type" content="text/html; charset=windows-1252"/>
  <title> Historia Apollonii regis Tyri </title>
  <link rel="SHORTCUT ICON" href="http://www.thelatinlibrary.com/icon.ico"/>
  <link rel="StyleSheet" href="http://www.thelatinlibrary.com/latinlibrary.css"/>
 </head>
 <body cz-shortcut-listen="true">
  <p class="pagehead">HISTORIA APOLLONII REGIS TYRI </p>
  <p class="border"></p>
  <p>
   <b>1</b> In civitate Antiochia rex fuit quidam nomine Antiochus, a quo ipsa civitas nomen accepit
   Antiochia. Is habuit unam filiam, virginem speciosissimam, in qua nihil rerum natura exerraverat,
   nisi quod mortalem statuerat. </p>
  <p> Quae dum ad nubilem pervenisset aetatem et species et formositas cresceret, multi eam in
   matrimonium petebant et cum magna dotis pollicitatione currebant. Et cum pater deliberaret, cui
   potissimum filiam suam in matrimonium daret, cogente iniqua cupididate flamma concupiscentiae
   incidit in amorem filiae suae et coepit eam aliter diligere quam patrem oportebat. Qui cum
   luctatur cum furore, pugnat cum dolore, vincitur amore; excidit illi pietas, oblitus est se esse
   patrem et induit coniugem. </p>
  <p> Sed cum sui pectoris vulnus ferre non posset, quadam die prima luce vigilans inrumpit
```

De “n'importe quoi” au texte brut

- Sources hétérogènes, codages hétérogènes
- Flux de travail:
  - Évaluation de chaque fichier / collection
  - Conversion de conventions typographiques et / ou de balisage en pseudo-balisage
  - Récupération de contenu textuel et conversion en texte brut
- Défis:
  - Conventions orthographiques (voyelles longues / courtes, u/v, i/j)
  - Abréviations
  - Informations éditoriales

Analyse linguistique automatique

- Entrée: fichiers texte brut
- Outil: librairie StanfordNLP pour Python (Stanza)
- Sortie: fichiers CONLL-U

Annotation automatique: code

```
import stanza
from stanza.utils.conll import CoNLL
nlp = stanza.Pipeline(lang="la", treebank="la_perseus")
input = open('source-text.txt', 'r')
doc = input.read()
results = nlp(doc)
annList = results.to_dict()
annConll = CoNLL.convert_dict(annList)
outf = "annotated-tex.txt"

with open(outf, mode="w") as f:
    for sent in annConll:
        for token in sent:
            print(*token, sep="\t", file=f)
        print("\n", file=f)
```

CONLL-U

```
1    quid    quis    PRON    p-s---na-    Case=Acc|Gender=Neut|Number=Sing    4    advmod    _    _
2    de    de    ADP    r--------    _    3    case    _    _
3    nobis    nos    PRON    p-p---mb-    Case=Abl|Gender=Masc|Number=Plur    4    obl    _    _
4    loquor    loquor    VERB    v1spip---    Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin|Voice=Pass    0    root    _    _
5    ?    ?    PUNCT    u--------    _    4    punct    _    _
```

CONLL-U

```
# Dale de comer al perro
1-2   Dale     _
1     Da       dar
2     le       él
3     de       de
4     comer    comer
5-6   al       _
5     a        a
6     el       el
7     perro    perro
```

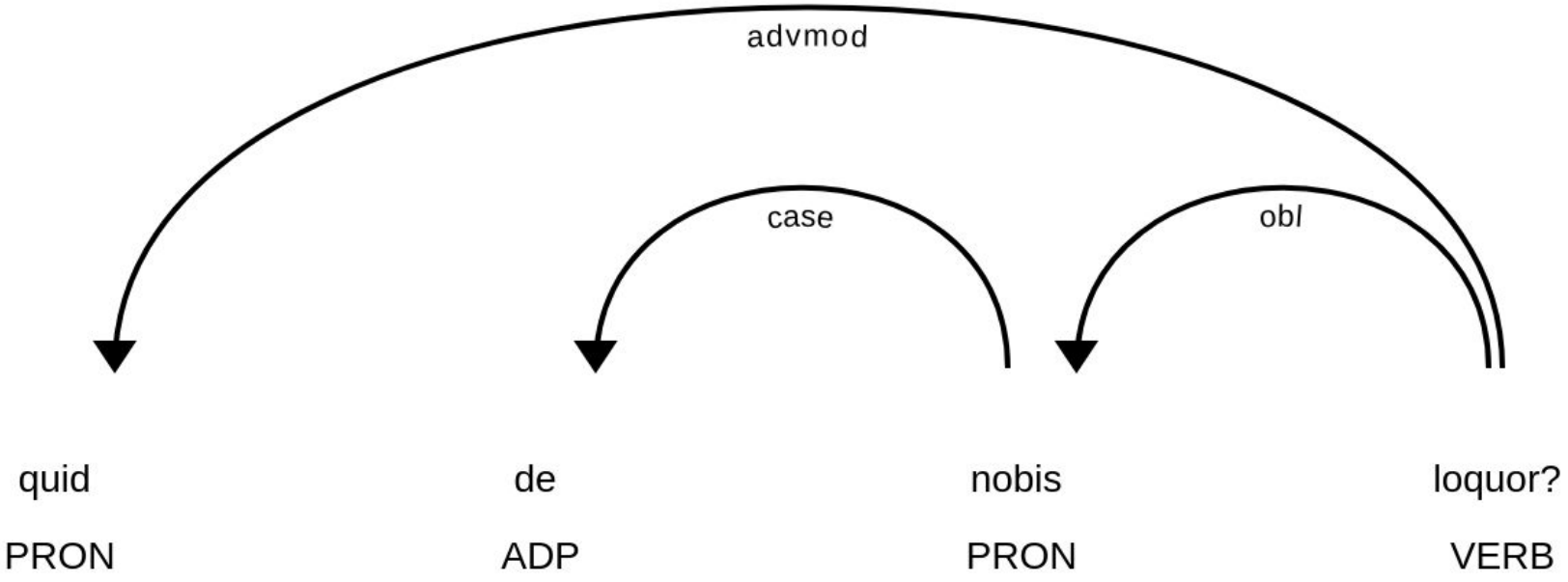
CONLL-U

```
# Sue likes coffee and Bill tea
1     Sue       Sue
2     likes     like
3     coffee    coffee
4     and       and
5     Bill      Bill
5.1   likes     like
6     tea       tea
```

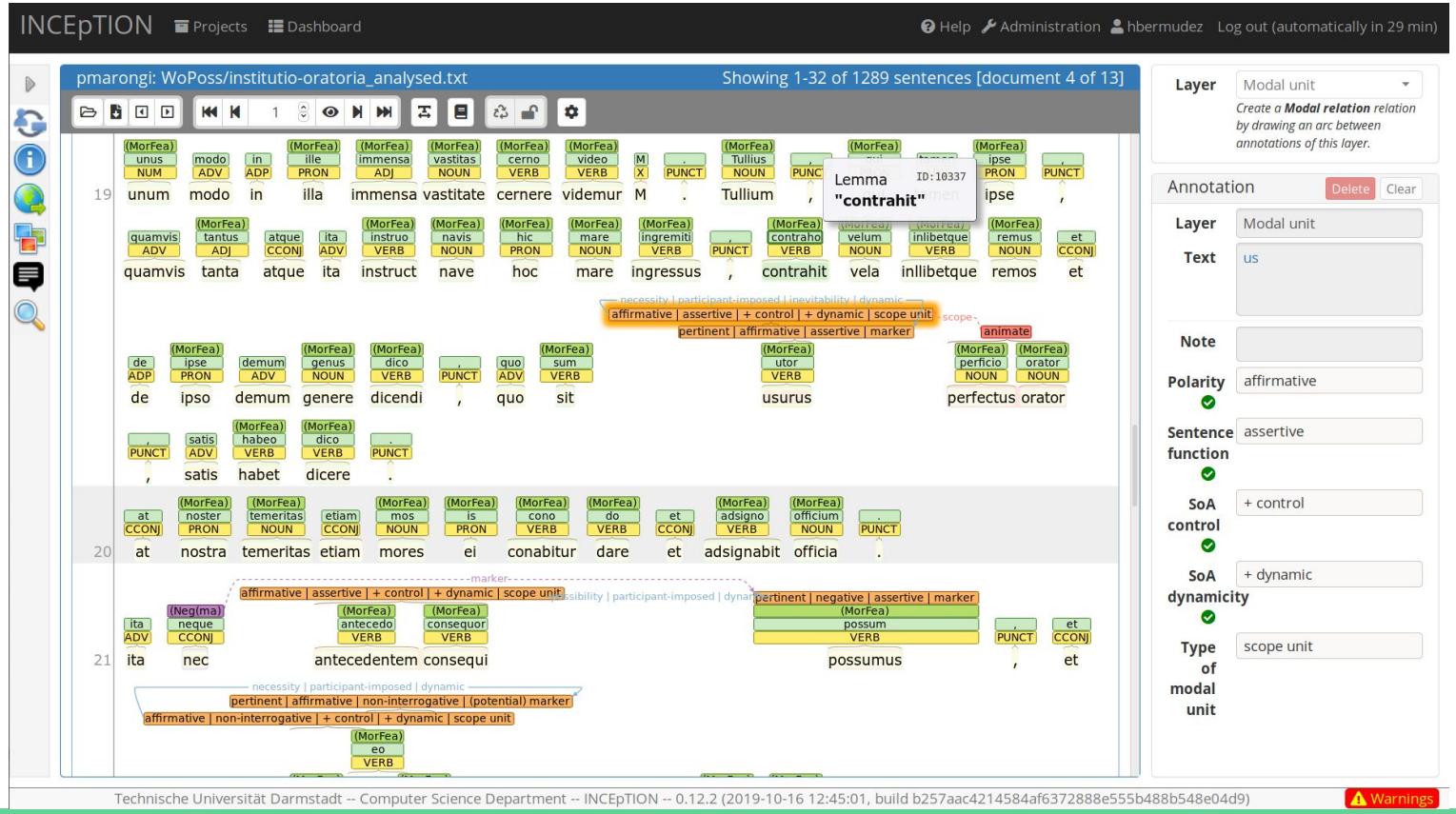
CONLL-U

```
1    quid    quis    PRON    p-s---na-    Case=Acc|Gender=Neut|Number=Sing    4    advmod    _    _
2    de    de    ADP    r--------    _    3    case    _    _
3    nobis    nos    PRON    p-p---mb-    Case=Abl|Gender=Masc|Number=Plur    4    obl    _    _
4    loquor    loquor    VERB    v1spip---    Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin|Voice=Pass    0    root    _    _
5    ?    ?    PUNCT    u--------    _    4    punct    _    _
```

CONLL-U



Annotation manuelle



Résultats de l'annotation manuelle

```
<type4:Token xmi:id="50442" sofa="1" begin="6674" end="6678" lemma="50461" pos="50455" order="0"/>
<type4:Token xmi:id="50474" sofa="1" begin="6679" end="6683" lemma="50493" pos="50487" order="0"/>
<type4:Token xmi:id="50506" sofa="1" begin="6684" end="6690" lemma="50548" pos="50542" morph="50519" order="0"/>
<type4:Token xmi:id="50561" sofa="1" begin="6691" end="6692" lemma="50580" pos="50574" order="0"/>
<type4:Token xmi:id="50598" sofa="1" begin="6693" end="6697" lemma="50640" pos="50634" morph="50611" order="0"/>
```

```
<morph:MorphologicalFeatures xmi:id="236540" sofa="1" begin="31161" end="31167" gender="Fem" number="Plur" case="Nom"
alue="Case=Nom|Gender=Fem|Number=Plur"/>
<morph:MorphologicalFeatures xmi:id="236595" sofa="1" begin="31168" end="31175" number="Plur" verbForm="Fin" tense="Pres" mood="Ind" voice="Pass"
alue="Mood=Ind|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin|Voice=Pass" person="1"/>
<morph:MorphologicalFeatures xmi:id="236682" sofa="1" begin="31178" end="31188" number="Plur" verbForm="Fin" tense="Pres" mood="Ind" voice="Pass"
alue="Mood=Ind|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin|Voice=Pass" person="1"/>
```

```
<custom:Modalunit xmi:id="23337" sofa="1" begin="3037" end="3049" Hasnostateofaffairs="false" Isnegative="false" Tokenisnotmodal="false"
Tokenisnotpertinent="false" Typeofmodalunit="scope unit" Typeofutterance="non-interrogative" Polarity="affirmative" SoAcontrol="+ control"
SoAdynamicity="- dynamic"/>
<custom:Modalunit xmi:id="345328" sofa="1" begin="45229" end="45240" Hasnostateofaffairs="false" Isnegative="false" Tokenisnotmodal="false"
Tokenisnotpertinent="false" Typeofmodalunit="scope unit" Typeofutterance="non-interrogative" Polarity="affirmative" SoAcontrol="+ control"
SoAdynamicity="+ dynamic"/>
<custom:Modalunit xmi:id="346342" sofa="1" begin="45338" end="45350" Hasnostateofaffairs="false" Isnegative="false" Tokenisnotmodal="false"
Tokenisnotpertinent="false" Typeofmodalunit="scope unit" Typeofutterance="non-interrogative" Polarity="affirmative" SoAcontrol="+ control"
SoAdynamicity="- dynamic"/>
```

```
<custom:ParticipantIsparticipantofLink xmi:id="354657" role="scope" target="354582"/>
<custom:ParticipantIsparticipantofLink xmi:id="354666" role="scope" target="140032"/>
```

Transformation en TEI

```
<s>
   <w pos="PRON" lemma="qui"> quibus </w>
   <w pos="PRON" lemma="ipse"> ipse </w>
   <w pos="VERB" lemma="aio"> ait </w>
   <w pos="ADV" lemma="numquid"> numquid </w>
   <w pos="VERB" lemma="possum" msd="d1e43969">
      <seg xml:id="d1e43977" type="marker">potestis</seg>
   </w>
   <seg xml:id="d1e43985" type="scope">
      <w pos="NOUN" lemma="filius" msd="d1e43987"> filios </w>
      <w pos="NOUN" lemma="sponsum" msd="d1e44002"> sponsi </w>
      <w pos="SCONJ" lemma="dum" rend="visible"> dum </w>
      <w pos="ADP" lemma="cum" rend="visible"> cum </w>
      <w pos="PRON" lemma="ille" msd="d1e44032"> illis </w>
      <w pos="VERB" lemma="sum" msd="d1e44045"> est </w>
      <w pos="NOUN" lemma="sponsus" msd="d1e44057"> sponsus </w>
      <w pos="VERB" lemma="facio" msd="d1e44069"> facere </w>
      <w pos="VERB" lemma="iejuno" msd="d1e44081"> ieiunare </w>
   </seg>
</s>
```

```
<fs type="marker" corresp="d1e43977">
   <f name="utterance">
      <symbol value="interrogative"/>
   </f>
   <f name="polarity">
      <symbol value="affirmative"/>
   </f>
   <f name="locus">
      <string>Lc 5, 34</string>
   </f>
</fs>
```

Correction et enrichissement

- Révision et intégration des informations
  - Détection d'éventuels problèmes de l'annotation
  - Problèmes textuels (problèmes philologiques)
  - Problèmes de reconnaissance de texte
  - Annotation de contenus implicites liés à la modalité
  - Ajout de métadonnées (DHTK, Picca & Egloff 2017)
- Transformation du pseudo-balisage en éléments TEI

Produit final (jeu de données)

Corpus diachronique dans le TEI standard qui contient:

- Le lemme, la partie du discours, les caractéristiques morphologiques, les dépendances syntaxiques (non corrigées)
- L’annotation sémantique: modalité,
- L’annotation structurelle (versets, paragraphes, sections)
- Le discours direct
- Le genre littéraire
- Le type de transmission textuelle
- L’auteur: origine, période, sexe
- Autres catégories de typologie textuelle (traduction, texte dialogué)

Produit final: publication

- Développement d'une interface utilisateur graphique pour exploiter le corpus

https://woposs.unine.ch/search

Références citées

Jannidis, Fotis, et Julia Flanders. 2013. « A concept of data modeling for the humanities ». In *Digital Humanities 2013: Conference Abstracts*, 237-39. Lincoln: Center for Digital Research in the Humanities.

Picca, Davide, et Mattia Egloff. 2017. « DHTK: The Digital Humanities ToolKit ». In *Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe II)*, édité par A. Adamou, E. Daga, et L. Isaken, 2014:81-86. CEUR Wokshop Proceedings. http://ceur-ws.org/Vol-2014/paper-09.pdf.

Sinclair, John. 2004. « Corpus and Text: Basic Principles ». In *Developing Linguistic Corpora: a Guide to Good Practice*, édité par Martin Wynne. AHDS Guides to Good Practice. AHDS. http://users.ox.ac.uk/~martinw/dlc/chapter1.htm.

Wallis, S. et Nelson G. 2001. « Knowledge discovery in grammatically analysed corpora ». *Data Mining and Knowledge Discovery,* 5: 307–340.

WoPoss. *A world of possibilities. Modal pathways over an extra-long period of time: the diachrony of modality in the Latin language*. https://woposs.unine.ch

Pour en savoir plus sur WoPoss

Bermúdez Sabel, Helena (press). « Digital tools for semantic annotation: the WoPoss use case». *Bulletin de linguistique et des sciences du langage*, 30. [Preprint].

Dell'Oro, Francesca. 2019. « WoPoss guidelines for annotation ». Zenodo. doi:10.5281/zenodo.3560950.

Dell'Oro, Francesca, Helena Bermúdez Sabel, et Paola Marongiu. 2020. « Implemented to Be Shared: The WoPoss Annotation of Semantic Modality in a Latin Diachronic Corpus ». In *Sharing the Experience: Workflows for the Digital Humanities. Proceedings of the DARIAH-CH Workshop 2019 (Neuchâtel)*. Neuchâtel, Switzerland: DARIAH-CAMPUS. doi:10.5281/zenodo.3739440.