

Text transcription and artificial intelligence

Issues & challenges

Mathieu Goux mathieu.goux@unicaen.fr

Thursday 16 December 2021 12:30-14:00 CET (Webinar)

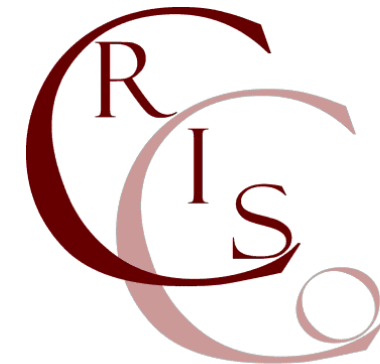
<https://www.disll.unipd.it/webinar-seminari-di-linguistica>



UNIVERSITÀ
DEGLI STUDI
DI PADOVA



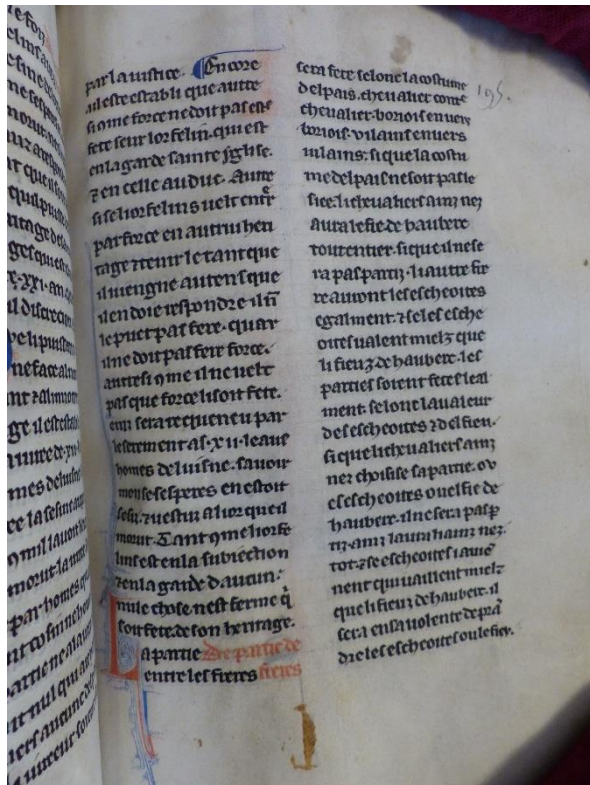
UNIVERSITÉ
CAEN
NORMANDIE



Centre de
Recherches
Inter-langues
sur la Signification
en COntexte

E.A. 4255

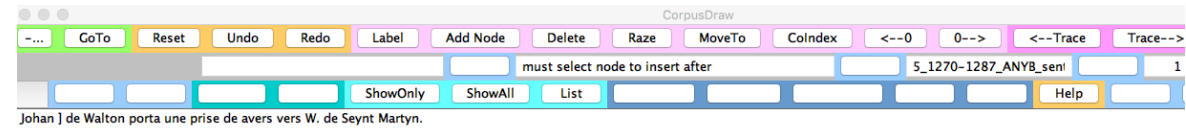
1. Corpus linguistics (CL) & NLP (Natural Language Processing)
2. Main challenges in long diachrony & in cross-linguistic research
3. Step by step workflow
4. References & Closing statements



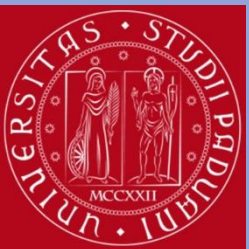
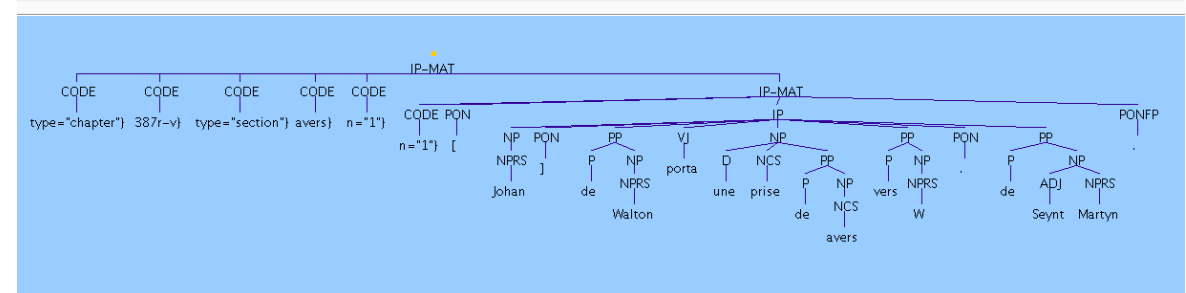
```

</ph>
<w n="19" pos="VJ">tent</w>
<ph function="IP-obj">
<w n="20" pos="D">cel</w>
<w n="21" pos="NCS">tenement</w>
<ph function="IP">
<ph function="IP-obl">
<w n="22" pos="WPRO">ou</w>
</ph>
<cl function="IP-SUB">
<ph function="IP-nsubj">
<w n="23" pos="D">la</w>
<w n="24" pos="NCS">prise</w>
</ph>
<w n="25" pos="VJ">fu</w>
</cl>
</ph>
</ph>

```



Johan] de Walton porta une prise de avers vers W. de Seynt Martyn.



1. Corpus linguistics, NLP & Deep learning

Corpus linguistics:

« The corpus is a fundamental tool for any type of research on language. The availability of computers in the 1950s immediately led to the creation of corpora in electronic form that could be searched automatically for a variety of language features, and compute frequency, distributional characteristics, and other descriptive statistics. » (Ide, 2008 : 328-329)

« ... Creation of linguistic corpora almost always demands that sub-paragraph structures such as sentences and words [...] are identified » (*ibid.* 332)

=> To interrogate and search *metadata* (aka, *data describing data*) and not directly *data*.

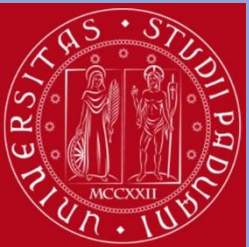
NLP & Deep Learning :

« At the June 2015 opening of the Facebook AI Research Lab in Paris, its director Yann LeCun said: “The next big step for Deep Learning is natural language understanding, which aims to give machines the power to understand not just individual words but entire sentences and paragraphs.” »

« Where has Deep Learning helped NLP? The gains so far have not so much been from true Deep Learning (use of a hierarchy of more abstract representations to promote generalization) as from the use of distributed word representations—through the use of real-valued vector representations of words and concepts. Having a dense, multidimensional representation of similarity between all words is incredibly useful in NLP, but not only in NLP. » (Manning, 2015 : 701, 703)

=> Deep Learning to help encode *PoS* & *syntactic* metadata

1. CL & NLP
2. Challenges
3. Workflow
4. Conclusion



=> Actually, we do not need to get data, but to describe them with metadata. As M. Eric put it (2021) « We don't need Data Scientists, we need Data Engineers » (<https://www.mihaileric.com/posts/we-need-data-engineers-not-data-scientists/>)

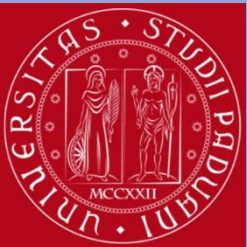
Stunning Data Growth Statistics (Editor's Choice)

It's now possible to see how much data is created every day, as well as how much data we consume regularly. You might be surprised to find out that:

- In 2020, people created 1.7 MB of data every second.
- By 2022, 70% of the globe's GDP will have undergone digitization.
- In 2021, 68% of Instagram users view photos from brands.
- By 2025, 200+ zettabytes of data will be in cloud storage around the globe.
- In 2020, users sent around 500,000 Tweets per day.
- By the end of 2020, **44 zettabytes** will make up the entire digital universe.
- Every day, **306.4 billion emails** are sent, and **500 million Tweets** are made.

<https://techjury.net/blog/how-much-data-is-created-every-day/#gref>

1. CL & NLP
2. Challenges
3. Workflow
4. Conclusion



⇒ As such, corpus linguistics & NLP / Deep Learning have several objectives:

- To *structure*, down to the word level, each text
- To *edit* the text in a universal format
- To *stabilize* annotation and metadata extraction
- To *open* the resources, for scientific collaboration
- To *develop* tools for data visualization and/or data structuration

⇒ The challenges NLP and CL face are completely different in nature:

⇒ Challenges in CL concern structural information, regarding the text as an entity.

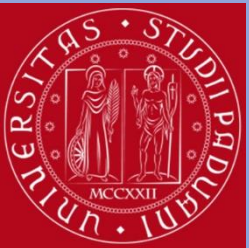
⇒ Challenges in NLP concern PoS & hierarchical information, regarding the text as a product of human language.

The answers are also quite distinct:

* challenges in CL can be answered automatically with a limited set of tools. For a computer, a textual data is no different in nature as any other kind of data (i.e. image, sounds, videos...), and features can be implemented manually;

* challenges in NLP can be answered automatically with a *greater* set of (+/- imperfect) tools, as for a computer, natural language cannot be understood (at all!), and features are hard to implement manually.

1. CL & NLP
2. Challenges
3. Workflow
4. Conclusion



2. Main challenges in long diachrony & in cross-linguistic research

2.1. Encoding

- By nature, CL & NLP aim to compare and run research on a great variety of texts
 - Inside a same language variety
 - Between several language varieties (i.e. diasystemic variation)
 - Between several languages

-> As such, CL & NLP imply some choices regarding textual data than can be hard to stabilize:

* PoS information: can we use the same PoS for each token without any problem?

Universal POS tags

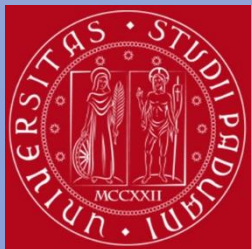
These tags mark the core part-of-speech categories. To dis [features](#).

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

- [Adjectifs \(ADJ\)](#)
- [Adverbes \(ADV\)](#)
- [Auxiliaires \(A, E, L\)](#)
- [Conjonctions \(CONJ\)](#)
- [Déterminants \(D\)](#)
- [À datif \(DAT\)](#)
- [Mots étrangers \(ETR\)](#)
- [Particules exprimant le focus \(FP\)](#)
- [Interjections \(ITJ\)](#)
- [Modaux \(MD\)](#)
- [Négation \(NEG\)](#)
- [Noms \(N\)](#)
- [Numéraux \(NUM\)](#)
- [Ponctuation \(PON\)](#)
- [Prépositions \(P\)](#)
- [Pronoms \(PRO\)](#)
- [Quantifieurs \(Q\)](#)
- [Verbes \(V\)](#)
- [Mots interrogatifs, relatifs et exclamatifs \(WADV, WD, WPRO\)](#)

<https://universaldependencies.org/u/pos/index.html>

<https://www.ling.upenn.edu/~beatrice/corpus-ling/annotation-french/pos/pos-index.html>

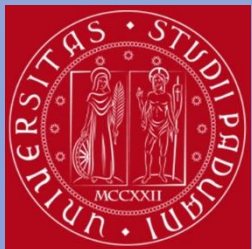


1. CL & NLP
2. Challenges
- 2.1. Encoding
- 2.2. Annotation
- 2.3. DataViz
3. Workflow
4. Conclusion

- Structural information: can we use the same syntactical informations for each language?

	Nominals	Clauses	Modifier words	Function Words
Core arguments	<u>nsubj</u> <u>obj</u> <u>iobj</u>	<u>csubj</u> <u>ccomp</u> <u>xcomp</u>		
Non-core dependents	<u>obl</u> <u>vocative</u> <u>expl</u> <u>dislocated</u>	<u>advcl</u>	<u>advmod</u> * <u>discourse</u>	<u>aux</u> <u>cop</u> <u>mark</u>
Nominal dependents	<u>nmod</u> <u>appos</u> <u>nummod</u>	<u>acl</u>	<u>amod</u>	<u>det</u> <u>clf</u> <u>case</u>
Coordination	MWE	Loose	Special	Other
<u>conj</u> <u>cc</u>	<u>fixed</u> <u>flat</u> <u>compound</u>	<u>list</u> <u>parataxis</u>	<u>orphan</u> <u>goeswith</u> <u>reparandum</u>	<u>punct</u> <u>root</u> <u>dep</u>

<https://universaldependencies.org/u/dep/index.html>



1. CL & NLP
2. Challenges
 - 2.1. Encoding
 - 2.2. Annotation
 - 2.3. DataViz
3. Workflow
4. Conclusion

- Which format should we use to encode our data?

```

</ph>
</ph>
<w n="19" pos="VJ">tent</w>
<ph function="IP-obj">
  <w n="20" pos="D">cel</w>
  <w n="21" pos="NCS">tenement</w>
  <ph function="IP">
    <ph function="IP-obl">
      <w n="22" pos="WPRO">ou</w>
    </ph>
    <cl function="IP-SUB">
      <ph function="IP-nsubj">
        <w n="23" pos="D">la</w>
        <w n="24" pos="NCS">prise</w>
      </ph>
      <w n="25" pos="VJ">fu</w>
    </cl>
  </ph>
</ph>

```

XML-TEI

```

#text notafter="1287" notbefore="1270" xml:id="ANYB_1270-1287" xml:lang="xno"
#body>
#div n="1" type="chapter">
#source>Cambridge University Library MS. Dd.7.14, ff. 387r-v#/source>
#div n="1" type="section">
#head>Prise de avers#/head>
#p n="1">
#s n="1">
#/s>
1  [ _ DET _ _ 2  obl _ _
2  Johan _ PROPN _ _ 6  nsubj _ _
3  ] _ ADV _ _ 2  appos _ _
4  de _ ADP _ _ 2  flat _ _
5  Walton _ PROPN _ _ 2  flat _ _
6  porta _ VERB _ _ 0  root _ _
7  une _ DET _ _ 8  det _ _
8  prise _ NOUN _ _ 6  obj _
9  de _ ADP _ _ 10 case _ _
10 avers _ NOUN _ _ 8  nmod _ _
11 vers _ ADP _ _ 12 case _ _
_ _ _ _ 8  nmod _ _
8  nmod _ _ _
13 case _ _ _
_ _ 16 amod _ _
_ _ 13 nmod _ _
13 advmod _ _

```

CONLLU

```

( (CODE n="1"))
( (IP-MAT (CODE n="1"))
  (PON [)
  (IP (NP (NPRS Johan))
    (PON ])
    (PP (P de)
      (NP (NPRS Walton)))
    (VJ porta)
    (NP (D une)
      (NCS prise)
      (PP (P de)
        (NP (NCS avers)))))
    (PP (P vers)
      (NP (NPRS W)))
    (PON .)
    (PP (P de)
      (NP (ADJ Seynt) (NPRS Martyn))))
  (PONFP .)))

```

PSD/PTB

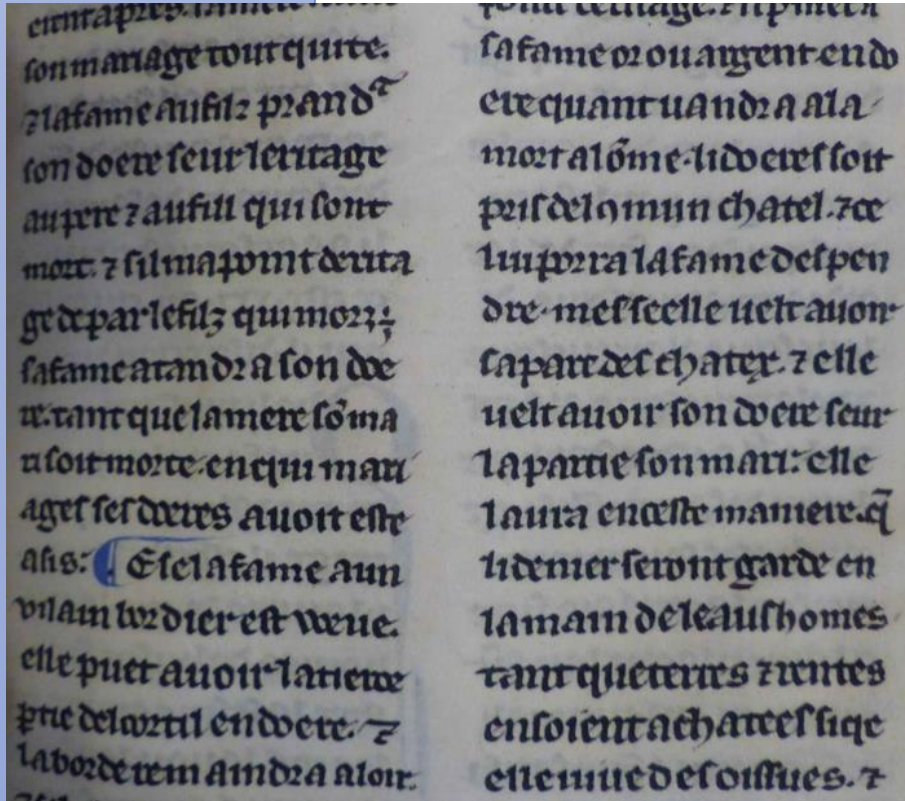
- CL & NLP
- Challenges

- 2.1. Encoding
- 2.2. Annotation
- 2.3. DataViz

3. Workflow
4. Conclusion

- But even before, how should we define a word/a token? A sentence? These concepts are quite recent and/or inadequate.

=> How to combine the specificity of one (variety of) language and universal descriptions?



Très Ancien Coutumier
(Norman Law, c. 1250)

- Latin: *Senatus Populus-que Romanus*

"Senate people-and Roman" = "The Senate and people of Rome"

- Ancient Greek: *ánthrōpoi (-te) theoi -te*

"people (and) gods and" = "(both) men and gods"

- Sanskrit: *naro gajaś '-ca* 'नरो गजश्च' i.e. "naraḥ gajaḥ ca" "नरस् गजस् -च" with sandhi

"the man the elephant and" = "the man and the elephant"

- Sanskrit: *Namaste* < *namaḥ + te*, (Devanagari: नमः + -ते = नमस्ते), with sandhi change *namaḥ > namas*.

"bowing to you"

http://www.digitorient.com/wp/wp-content/uploads/2015/04/Paleo-bab_complet.pdf

KUM (GUM)			kum, qum, gum, qu	kum
GAZ (GUMxŠE)				gaz
ÚR			kàs	
IL			il	il

1. CL & NLP
2. Challenges

- 2.1. Encoding
- 2.2. Annotation
- 2.3. DataViz

3. Workflow
4. Conclusion

- To sum up:
 - Standardization challenges:
 - * metalinguistic units (sentences/tokens)
 - * metalinguistic tags (PoS/Syntax/lemma)
 - * format standardization (XML-TEI/CONLL/PSD-PTB)
 - ...and skills training! Community tools and digital humanities procedures already exist, but we have to know how to use them.

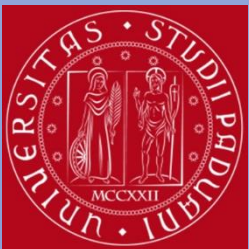
1. CL & NLP
2. Challenges

2.1. Encoding
2.2. Annotation
2.3. DataViz

3. Workflow
4. Conclusion

User	Edit	Context	Old Lemma	Corr Lemma	Previous POS	Actual POS	Previous Morph	Actual Morph	Similar
S.Gabay	Feb 11, 2019 4:55 PM	D' agréables langueurs et de ravissements , Jusques où d' un bel oeil peut s'	jusque	jusque	VERc j g	PRE	-	-	0 similar
S.Gabay	Feb 11, 2019 4:55 PM	saisissements , D' agréables langueurs et de ravissements , Jusques où d' un bel oeil	ravissement	ravissement	VERc j g	NOMcom	-	-	0 similar
S.Gabay	Feb 11, 2019 4:55 PM	appas , de doux saisissements , D' agréables langueurs et de ravissements , Jusques où	agréable	agréable	NOMcom	ADJqua	-	-	0 similar
S.Gabay	Feb 11, 2019 4:55 PM	de sorte d' appas , de doux saisissements , D' agréables langueurs et de ravissements	saisissement	saisissement	VERc j g	NOMcom	-	-	0 similar
S.Gabay	Feb 11, 2019 4:54 PM	amour a des tendresses Que nous n' apprenons point qu' auprès de nos maîtresses .	apprendre	apprendre	NOMcom	VERc j g	-	-	0 similar
S.Gabay	Feb 11, 2019 4:54 PM	faut point douter , l' amour a des tendresses Que nous n' apprenons point qu'	de_le	un	DETndf	DETndf	-	-	82 similar
S.Gabay	Feb 11, 2019 4:53 PM	d' un amant parfait . Il n' en faut point douter , l' amour a	en	en	PRE	PROadv	-	-	334 similar
S.Gabay	Feb 11, 2019 4:53 PM	l' avoir bien fait ; Un bon poète ne vient que d' un amant parfait	poète	poète	VERc j g	NOMcom	-	-	0 similar
S.Gabay	Feb 11, 2019 4:53 PM	discourir , il faut l' avoir bien fait ; Un bon poète ne vient que	faire	faire	VERc j g	VERppe	-	-	76 similar
S.Gabay	Feb 11, 2019 4:53 PM	Pour en bien discourir , il faut l' avoir bien fait ; Un bon poète	le	le	DETdef	PROper	-	-	435 similar

<https://pyrrha.readthedocs.io/en/latest/>



2.2. Annotation

- Even if we all agree on a PoS/syntactic/token standard, we have to annotate the texts. There are two ways to do it:
 - Statically, through the comparison of the texts with a dictionary and patterns recognition;
 - Dynamically, through a treebank and NPL/DL models.
 - A *static* approach helps us to have a greater control of the output, but manual corrections are to be expected;
 - A *dynamic* approach gives us less control of the output, but corrections are scarce.
- => Of course, a dynamic approach can only be adopted if a training model exists.

1. CL & NLP
2. Challenges

- 2.1. Encoding
- 2.2. Annotation
- 2.3. DataViz

3. Workflow
4. Conclusion

Accueil Documentation Interroger Télécharger Travaux Utilisateurs

Un corpus de référence pour le français

Une ressource lexicale et syntaxique richement annotée (et validée manuellement) pour les utilisables en TAL.

- Projet initié en 1997, avec le soutien de l'UIF, du CNRS et du CNRTL
- 21 550 phrases (environ 664 500 tokens) du journal *Le Monde* (1990-1993)
- métadonnées : auteur, date, domaine (par article)
- Annotations lexicales (catégories, sous-catégories, flexion, mots composés avec composants) et s (constituants majeurs, fonctions grammaticales) validées
- [Corpus annoté téléchargeable](#) (version 1.0 2016) en plusieurs formats (xml, Tiger-xml, PTB, CoNLL)

Le corpus arboré est diffusé gratuitement à des fins de recherche, sous réserve de la signature des [d'utilisation](#)

✓ [Nous contacter](#) pour obtenir une licence permettant une utilisation commerciale et le développement de dérivés

Portuguese(100)

[context] [conllu]



http://depsearch-depsearch.rahtiapp.fi/ds_demo/

- * Necessary when there is no treebank or model accessible (i.e. Old Italian/ Venetian)
- * Necessary when working on peculiar texts (low level of literacy, marginal annotations, drafts...)

L'ÉCRITURE DES PEU LETTRÉS : FRANÇAIS VERNACULAIRE DANS LA NORMANDIE MÉDIÉVALE

Le projet EPELE a pour objectif de réunir des éléments indicatifs de la pratique vernaculaire du français à l'époque médiévale. Productions transitoires de la part de lettrés, à réunis nous informent sur l'écriture, la prononciation et la grammaire Normandie à cause de l'importance qu'y a la production de l'écrit, pour but d'encourager des initiatives comparables pour une meilleure



The Prize Papers Project Events Pri

<https://www.unicaen.fr/epele/accueil>



<https://www.prizepapers.de/>

1. CL & NLP
2. Challenges
 - 2.1. Encoding
 - 2.2. Annotation
 - 2.3. DataViz
3. Workflow
4. Conclusion

Dynamic approach:

- * Can parse and annotate a great volume of texts (2 mn for 50 000+ tokens in Old French with the HOPS model)
- * Often trained on specific texts (press, littérature... see Camps *et al.* 2020), that can be problematic.

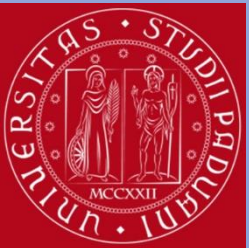
« If many lemmatisers and POS taggers have been trained, and sometimes conceived, for French, they usually focus on contemporary French and tools for Ancien Régime French remain scarce. [...] training data are not publicly available (yet?), and rely mainly on non-normalised texts from the 16th to the 18th c. » (Camps *et al.*, 2020 : 1)

=> In reality, as manual corrections are inevitable, a *mixed* approach should be considered.

=> When we look around, we see that projects and annotation system, dictionaries, data... are numerous but generally, it is hard to make those projects communicate to each other.

=> We should aim to add as much information (i.e. metadata/documentation) as possible.

1. CL & NLP
2. Challenges
 - 2.1. Encoding
 - 2.2. Annotation
 - 2.3. DataViz
3. Workflow
4. Conclusion




```

1 [Johan] _ PROPN _ _ 4 nsubj _ _
2 de _ ADP _ _ 1 flat _ _
3 Walton _ PROPN _ _ 1 flat _ _
4 porta _ VERB _ _ 0 root _ _
5 une _ DET _ _ 6 det _ _
6 prise _ NOUN _ _ 4 obj _ _
7 de _ ADP _ _ 8 case _ _
8 avers _ NOUN _ _ 6 nmod _ _
9 vers _ ADP _ _ 10 case _ _
10 W. _ PROPN _ _ 4 obl _ _
11 de _ ADP _ _ 13 case _ _
12 Seynt _ ADJ _ _ 13 amod _ _
13 Martyn _ PROPN _ _ 10 nmod _ _
14 . _ ADV _ _ 10 advmod _ _

```

```

1 [Johan] [johan] PROPN _ _ 4 nsubj _ _
2 de un ADP _ _ 1 flat _ PrestoPOS=Dn
3 Walton walton PROPN _ _ 1 flat _ _
4 porta porter VERB _ _ 0 root _ PrestoPOS=Vvc
5 une une DET _ _ 6 det _ PrestoPOS=Nc
6 prise priser NOUN _ _ 4 obj _ PrestoPOS=Vvc
7 de un ADP _ _ 8 case _ PrestoPOS=Dn
8 avers avers NOUN _ _ 6 nmod _ PrestoPOS=Nc
9 vers vert ADP _ _ 10 case _ PrestoPOS=Nc
10 W. w. PROPN _ _ 4 obl _ _
11 de un ADP _ _ 13 case _ PrestoPOS=Dn
12 Seynt _ ADJ _ _ 13 amod _ _
13 Martyn martyn PROPN _ _ 10 nmod _ _
14 . . ADV _ _ 10 advmod _ PrestoPOS=Fs
#/s>

```

```

couloyer/VER/Vvc/COULER/COULER/INC
couloyes/VER/Vvc/COULER/COULER/INC
couloyez/VER/Vvc/COULER/COULER/INC
couloyr/NOM/Nc/COULOIR/COULOIR/INC
couloyre/NOM/Nc/COULOIRE/COULOIRE/INC
couloyres/NOM/Nc/COULOIRE/COULOIRE/INC
couloyrez/NOM/Nc/COULOIRE/COULOIRE/INC
couloyrs/NOM/Nc/COULOIR/COULOIR/INC
couloyrz/NOM/Nc/COULOIR/COULOIR/INC
couloyré/NOM/Nc/COULOIRE/COULOIRE/INC
couloys/VER/Vvc/COULER/COULER/INC
couloyt/VER/Vvc/COULER/COULER/INC
couloyz/VER/Vvc/COULER/COULER/INC
couloyés/VER/Vvc/COULER/COULER/INC
couloz/NOM/Nc/COULOT/COULOT/INC
coupable/ADJ/Ag/COUPABLE/COUPABLE/INC
coupable/NOM/Nc/COUPABLE/COUPABLE/INC
coupables/ADJ/Ag/COUPABLE/COUPABLE/INC
coupables/NOM/Nc/COUPABLE/COUPABLE/INC
coupablez/ADJ/Ag/COUPABLE/COUPABLE/INC
coupablez/NOM/Nc/COUPABLE/COUPABLE/INC
coupans/PAG/Ga/COULPER/COULPER/INC
coupant/PAG/Ga/COULPER/COULPER/INC
coupante/PAG/Ga/COULPER/COULPER/INC

```

HOPS => PRESTO dictionary => "upgraded" CONLLU
(courtesy to the HIGH-TECH Project)

1. CL & NLP
2. Challenges

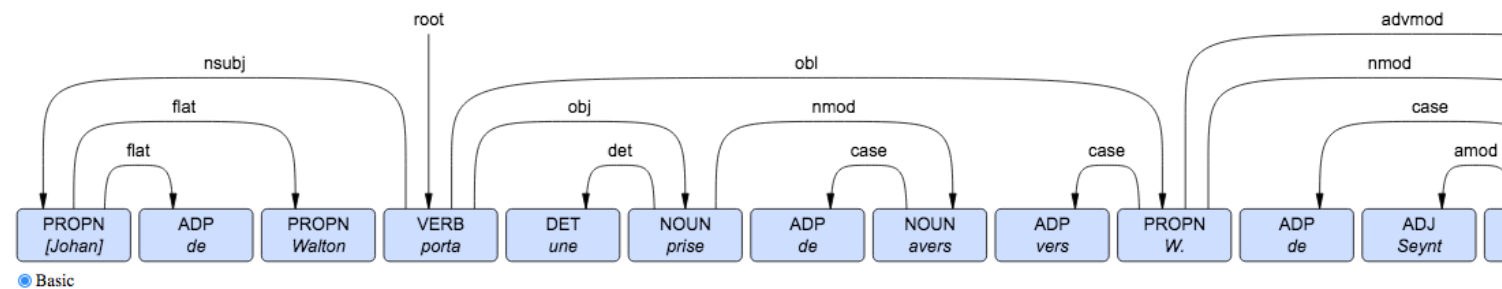
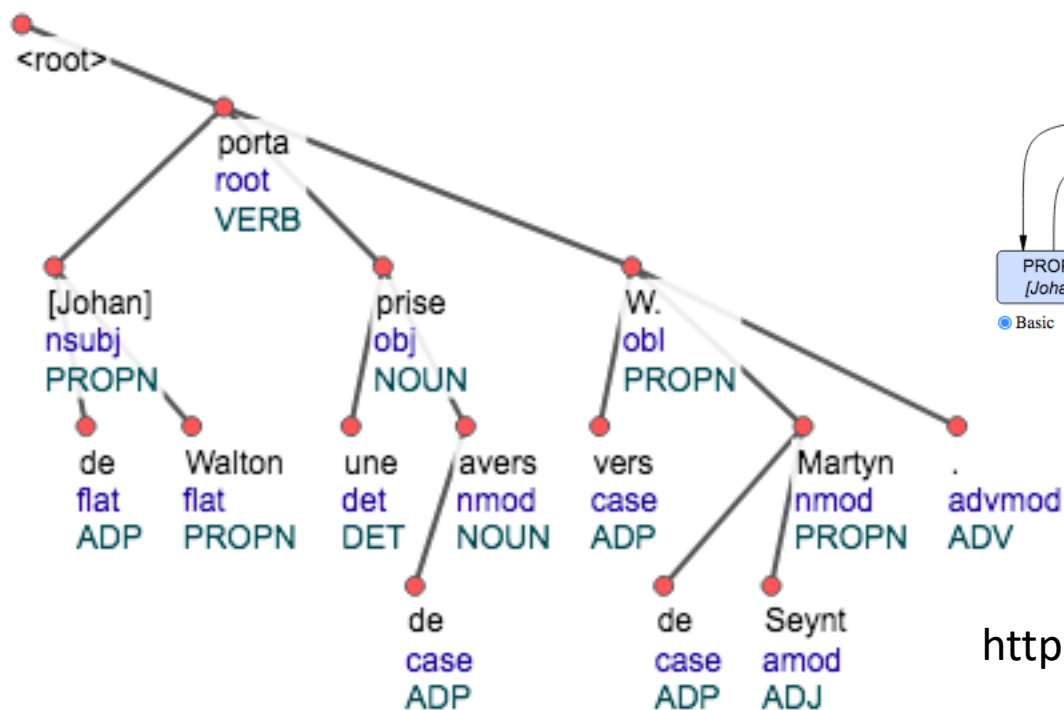
- 2.1. Encoding
- 2.2. Annotation
- 2.3. DataViz

3. Workflow
4. Conclusion

2.3. Visualization

- But, even if everything is standardized and « perfect », there is one last problem: data visualization.
- Basically, there are two ways to see data:
 - Graphically, i.e a « tree »;
 - In a table format, i.e. a csv.
- But even here, there are different ways to draw a tree:

[Johan] de Walton porta une prise de avers vers W. de Seynt Martyn .



<https://urd2.let.rug.nl/~kleiweg/conllu/>

https://universaldependencies.org/conllu_viewer.html

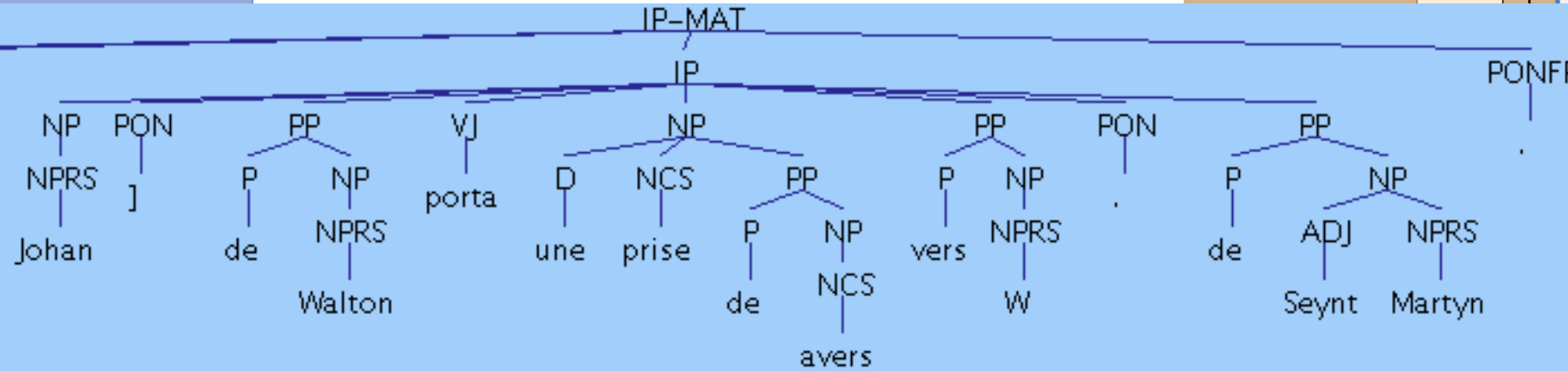
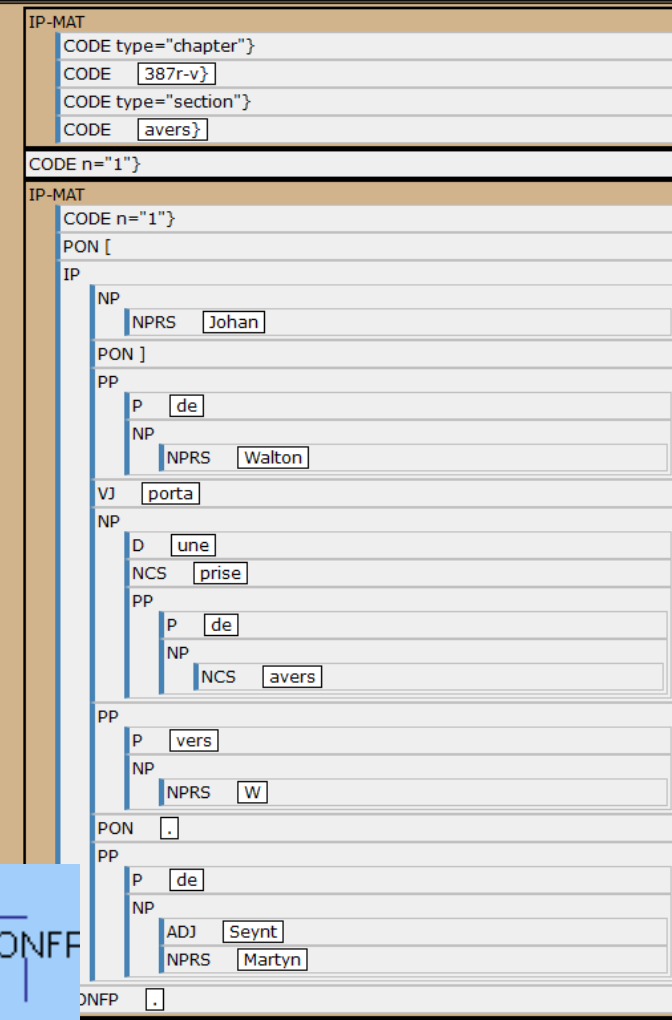
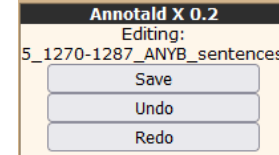
- More visualizations!

```
1 sentence.print_tree()
```

```
(deprel:root) form:porta lemma:_ upos:VERB [6]
  (deprel:nsubj) form:Johan lemma:_ upos:PROPN [2]
    (deprel:obl) form:[ lemma:_ upos:DET [1]
      (deprel:appos) form:] lemma:_ upos:ADV [3]
      (deprel:flat) form:de lemma:_ upos:ADP [4]
      (deprel:flat) form:Walton lemma:_ upos:PROPN [5]
    (deprel:obj) form:prise lemma:_ upos:NOUN [8]
      (deprel:det) form:une lemma:_ upos:DET [7]
      (deprel:nmod) form:avers lemma:_ upos:NOUN [10]
        (deprel:case) form:de lemma:_ upos:ADP [9]
      (deprel:nmod) form:W lemma:_ upos:PROPN [12]
        (deprel:case) form:vers lemma:_ upos:ADP [11]
      (deprel:nmod) form:. lemma:_ upos:ADV [13]
        (deprel:case) form:de lemma:_ upos:ADP [14]
        (deprel:nmod) form:Martyn lemma:_ upos:PROPN [16]
          (deprel:amod) form:Seynt lemma:_ upos:ADJ [15]
        (deprel:advmod) form:. lemma:_ upos:ADV [17]
```

<https://pypi.org/project/conllu/>

[annotald.github.io/](https://github.com/annotald/annotald)



corpussearch.sourceforge.net/

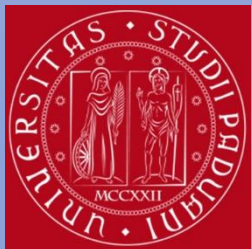
- A .csv is a highly standardized format, than can be readable and searchable using different tools. It is also quite useful to spot and compare patterns between several texts, as we search for metadata.

Requête : [suppos="P"] Chercher

	Référence	Contexte gauche	Pivot	Contexte droit
328	7_1406_Gauvard	pour un ouir dire si en termes generaux	sans	autre coulour on deveroit proposer un tel cas si redundant ou prejudice
329	7_1406_Gauvard	les abilemens que on dit avoir esté trouvez	en	la possession des Haies sont presumpcions que ce n'est pas espieurs
330	7_1406_Gauvard	on dit avoir esté trouvez en la possession	des	Haies sont presumpcions que ce n'est pas espieurs de chemins sed
331	7_1406_Gauvard	sont presumpcions que ce n'est pas espieurs	de	chemins sed pocius fines nocturni. Dit que de la faveur bien appert
332	7_1406_Gauvard	de chemins sed pocius fines nocturni. Dit que	de	la faveur bien appert car l'un qui estoit clerc a esté
333	7_1406_Gauvard	qu'ilz ont rez, noté que ce feust	pour	la tonsure que il avoit par avant car l'autre qui aussi
334	7_1406_Gauvard	ce feust pour la tonsure que il avoit	par	avant car l'autre qui aussi fut pendu ne fut point rez
335	7_1406_Gauvard	qui aussi fut pendu ne fut point rez	pour	ce qu'il n'avoit point de coronne. Quant a Hannotin qui
336	7_1406_Gauvard	rez pour ce qu'il n'avoit point	de	coronne. Quant a Hannotin qui fut rendu dit que des Hayes a
337	7_1406_Gauvard	qu'il n'avoit point de coronne. Quant	a	Hannotin qui fut rendu dit que des Hayes a la mort comme
338	7_1406_Gauvard	Quant a Hannotin qui fut rendu dit que	des	Hayes a la mort comme on dit le descoulpa. Conclut que il
339	7_1406_Gauvard	Hannotin qui fut rendu dit que des Hayes	a	la mort comme on dit le descoulpa. Conclut que il fait a
340	7_1406_Gauvard	on dit le descoulpa. Conclut que il fait	a	recevoir et alias ut supra et que on doit adjouster foy au
341	7_1406_Gauvard	court verra tout ce que les parties vaudront	en	conseil et en arrest.
342	7_1406_Gauvard	ce que les parties vaudront en conseil et	en	arrest.
343	7_Murano_custom		A@	@ciò che li beni de@ @le ghiesie de questa terra de Muran
344	7_Murano_custom		A@	@ciò che li beni de@ @le ghiesie de questa terra de Muran, siano ben custodidi et
345	7_Murano_custom	A@ @ciò che li beni de@ @le ghiesie	de	questa terra de Muran, siano ben custodidi et non vadino in
346	7_Murano_custom	li beni de@ @le ghiesie de questa terra	de	Muran, siano ben custodidi et non vadino in sinistro, però
347	7_Murano_custom	Muran, siano ben custodidi et non vadino	in	sinistro, però sia statuido et ordenado che per cadauna ghiesia parochial
348	7_Murano_custom	sinistro, però sia statuido et ordenado che	per	cadauna ghiesia parochial de questa terra siano electi do procuratori, citadini
349	7_Murano_custom	statuido et ordenado che per cadauna ghiesia parochial	de	questa terra siano electi do procuratori, citadini de Muran per anni
350	7_Murano_custom	questa terra siano electi do procuratori, citadini	de	Muran per anni do per miser lo piovàn et sacerdoti de@ @la
351	7_Murano_custom	siano electi do procuratori, citadini de Muran	per	anni do per miser lo piovàn et sacerdoti de@ @la ghiesia.

txm-crisco.huma-num.fr/

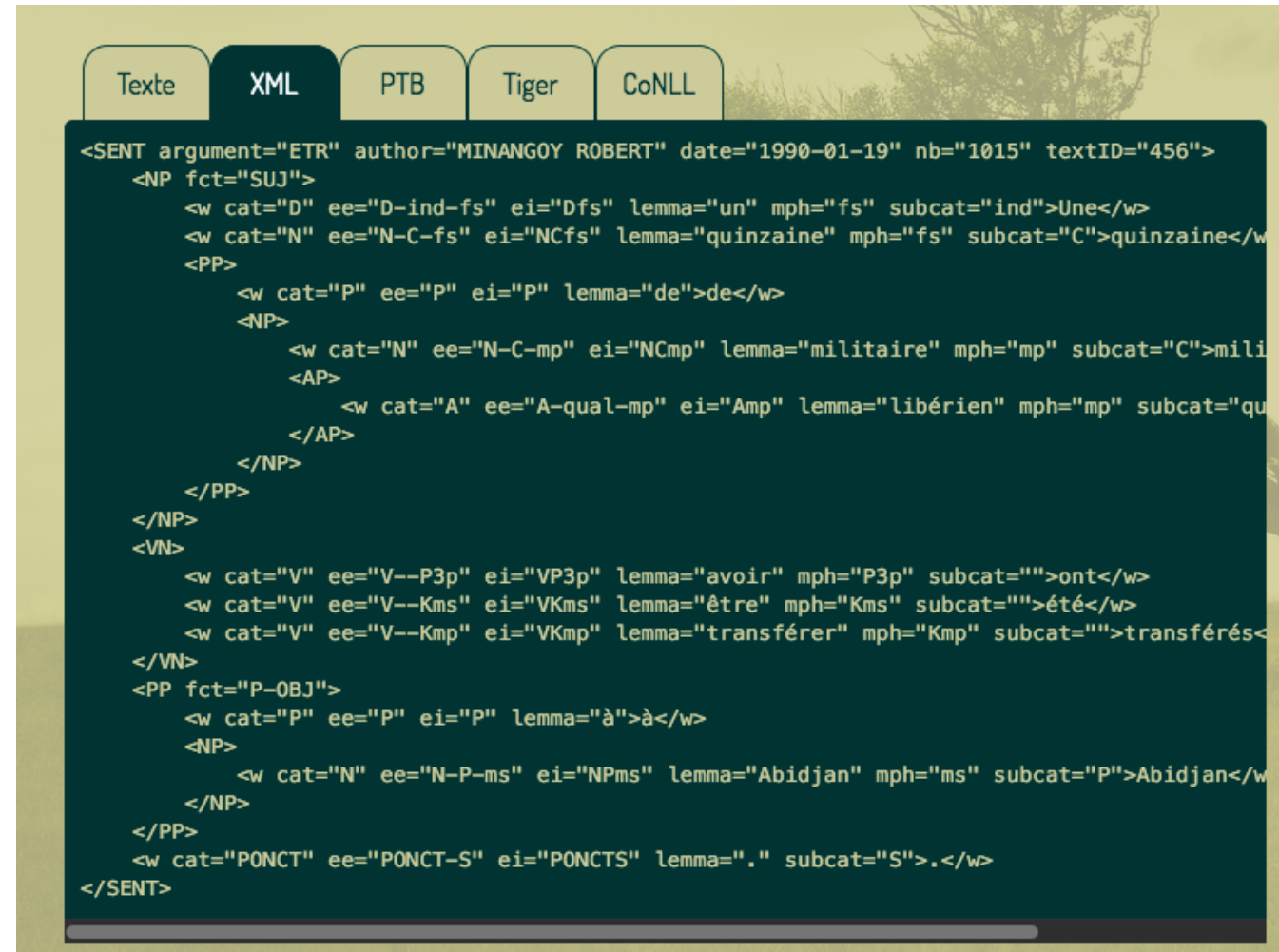
=> Again, we should aim to produce « universal data », that can be parsed and visualized using different tools, as the endgoal of CL/NLP is not to *create* data, but to *explore* it.



1. CL & NLP
2. Challenges
- 2.1. Encoding
- 2.2. Annotation
- 2.3. DataViz
3. Workflow
4. Conclusion

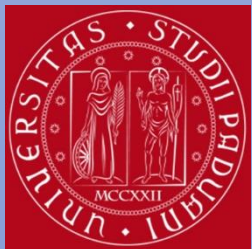
- ⇒ It is quite straightforward to convert from one format to another (e.g., CONLL ↔ XML-TEI ↔ PSD). XML-TEI is, however, the most practical and should be the « endgame » of every project, as it
- facilitates visualization
 - Can enable access to the texts via a webpage (with a XSLT / HTML transformation)

1. CL & NLP
2. Challenges
 - 2.1. Encoding
 - 2.2. Annotation
 - 2.3. DataViz
3. Workflow
4. Conclusion



```
<SENT argument="ETR" author="MINANGOY ROBERT" date="1990-01-19" nb="1015" textID="456">
  <NP fct="SUJ">
    <w cat="D" ee="D-ind-fs" ei="Dfs" lemma="un" mph="fs" subcat="ind">Une</w>
    <w cat="N" ee="N-C-fs" ei="NCfs" lemma="quinzaine" mph="fs" subcat="C">quinzaine</w>
    <PP>
      <w cat="P" ee="P" ei="P" lemma="de">de</w>
      <NP>
        <w cat="N" ee="N-C-mp" ei="NCmp" lemma="militaire" mph="mp" subcat="C">mili
        <AP>
          <w cat="A" ee="A-qual-mp" ei="Amp" lemma="libérien" mph="mp" subcat="qu
          </AP>
        </NP>
      </PP>
    </NP>
    <VN>
      <w cat="V" ee="V--P3p" ei="VP3p" lemma="avoir" mph="P3p" subcat="">ont</w>
      <w cat="V" ee="V--Kms" ei="VKms" lemma="être" mph="Kms" subcat="">été</w>
      <w cat="V" ee="V--Kmp" ei="VKmp" lemma="transférer" mph="Kmp" subcat="">transférés<
    </VN>
    <PP fct="P-OBJ">
      <w cat="P" ee="P" ei="P" lemma="à">à</w>
      <NP>
        <w cat="N" ee="N-P-ms" ei="NPms" lemma="Abidjan" mph="ms" subcat="P">Abidjan</w>
      </NP>
    </PP>
    <w cat="PONCT" ee="PONCT-S" ei="PONCTS" lemma="." subcat="S">.</w>
  </SENT>
```

<http://ftb.linguist.univ-paris-diderot.fr/#tag>



3. Step by step workflow

Three main steps (https://www.unicaen.fr/projet_de_recherche/micle/)

- a. Data acquisition
- b. Data structuring
- c. Data visualization

/!\ Each step requires several (philological/encoding/syntactic...) choices. Depending on the nature of the texts, all projects do not need to make the same choices to be interoperable, *but* they need to share a common language. The TEI-C aims, since 1994, to define a stable set of tags that can be used for textual data. /!\

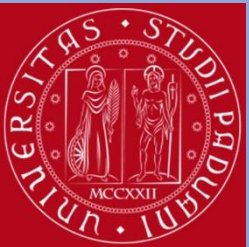


Text Encoding Initiative

The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of text. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences, and languages. Since 1994, the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for teaching, and preservation. In addition to the Guidelines themselves, the Consortium provides a variety of [resources](#) and [training](#) information on [projects using the TEI](#), a [bibliography of TEI-related publications](#), and [software](#) developed for or adapted to the TEI.

<https://tei-c.org/>

1. CL & NLP
2. Challenges
3. Workflow
 - 3.1. Acquisition
 - 3.2. Structuring
 - 3.3. DataViz
4. Conclusion



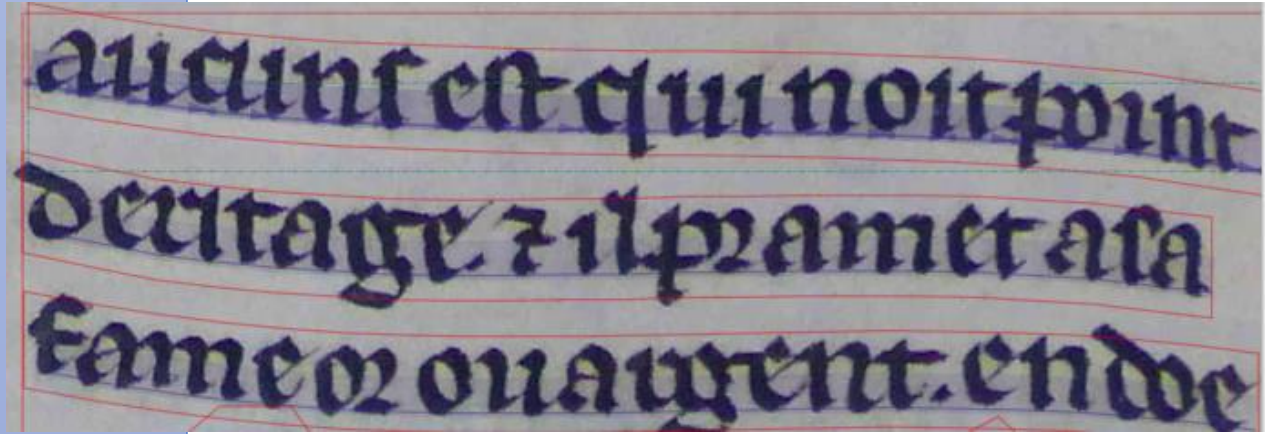
3.1 Data acquisition

- The end goal in this step is to produce a plain txt text or a set of XML-TEI files with information on the basic structure of the text.
- As such, challenges & issues arise:
 - If we are working on a manuscript, do we keep the characters, the line breaks?
 - If we are working on a printed text, do we keep page numbers and running titles?

=> Many projects in CL seek not only to interrogate the texts, but also to give access to the philological matter, to allow comparisons between editions.

<https://www.rialfri.eu/rialfriWP/>

I		LassaVerso	Ed. Thomas 1913	Ed. Infurna 2011	
	En honor et en bien et en gran remembrance				
	Et offerant mercé, honor et celebrance	2	25	tenser	saucer
	De Celui che par nos fu feruq de la lance	3	48	scrist	scrit
	Par trer nos e nos armes de la enfernal poiss	6	96	qe	que
5	Et de son saint apostre, qi tant oit penetance	7	139	qi	qui
	Por feir qe cescuns fust en veraie creance	10	241	noz	nos
	Que Per e Filz e Spirt sunt in une sustance		244	obli a	oblia



1-27 la tierce part en doere. C Se

2-1 aucunf est qui noit point

2-2 deritage. 7 il pramet a fa

2-3 fame o2 ou argent. en doe

```
<Baseline points="1574,715 1626,723 1679,730 1732,715">
<TextEquiv>
  <Unicode>re. C Se aucunf est q noit</Unicode>
</TextEquiv>
</TextLine>
<TextLine id="r215" custom="readingOrder {index:4;}">
```

```
8 fon [?]oere feur leritage
9 au pere 7 au fill qui font
10 mort. 7 fil ni a point [?]erita
11 ge [?]e par le filz qui morz ;
12 fa fame atan[?]ra fon [?]oe
13 re. tant que la mere fō ma
14 ri foit morte. en qui mari
15 agef fef [?]oeres auoit este
16 afis: C E fe la fame a un
17 vilain boz[?]ier est vveue.
18 elle puet auoir la tierce
19 ptie [?]el cortil en [?]oere. 7
20 la boz[?]le remain[?]ra a loir.
```

Junicode

A new version of Junicode is under development at <https://github.com/readcoop/junicode> hosted here is suspended. Minor enhancements and fixes for the leg these will be released in the "legacy" folder of the new Junicode site [/legacy](#).

Junicode (short for Junius-Unicode) is a TrueType/OpenType font for various ranges, plus Runic and Gothic. The font comes in four faces. Of the implementation of the [Medieval Unicode Font Initiative](#) recommend

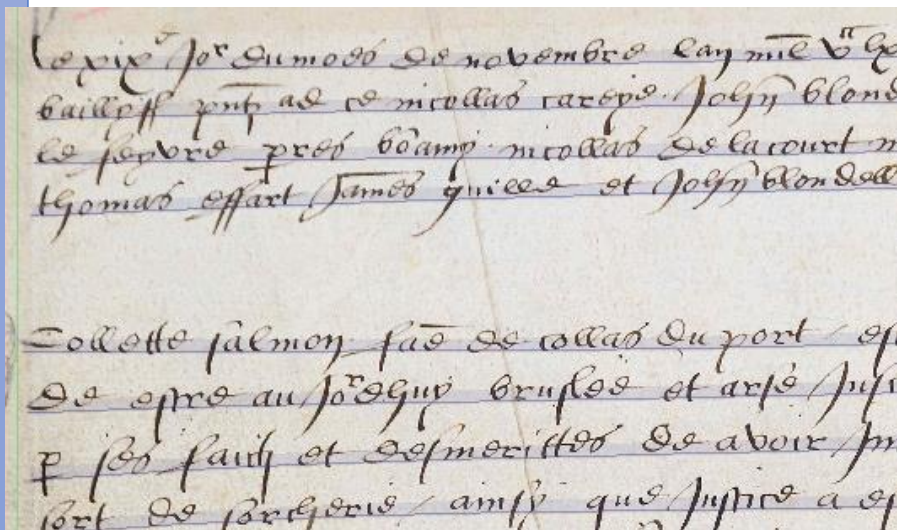
```
</w>
'><choice><orig>fef</orig><reg>ses</reg></choice></w>
724">uo<choice><orig>1f</orig><reg>is</reg></choice>ins</w>
```

- ⇒ Challenges in data acquisition:
 - ⇒ What is a character/a line/a text?
 - ⇒ Do we keep facsimile information (iiif)?
 - ⇒ Do we keep track of the different editions?
 - ⇒ How can we access the data?

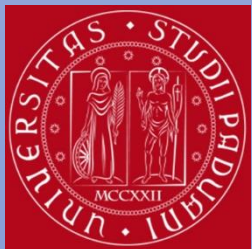
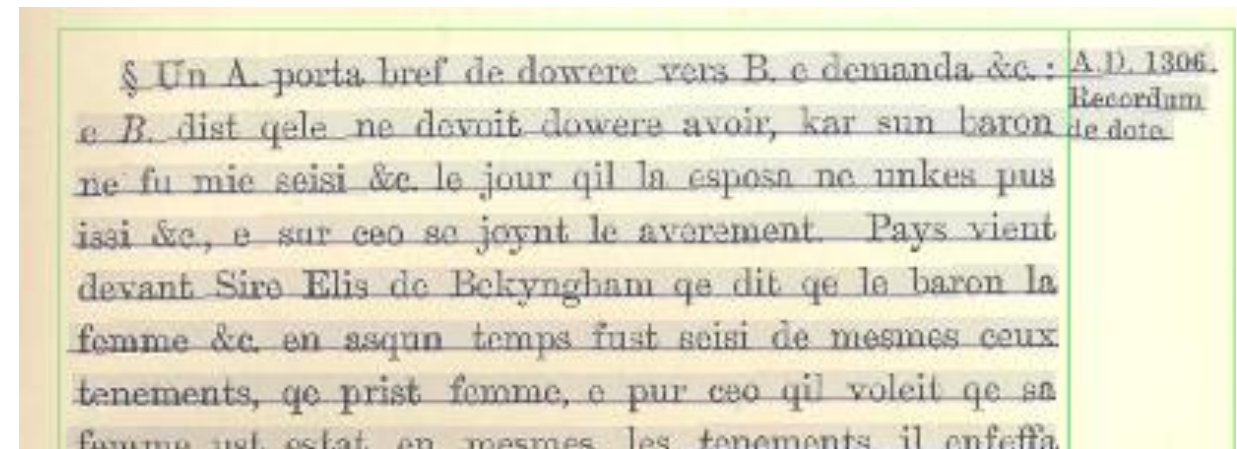
- Even if our only concern is the linguistic data, concepts like mediality and affordances can influence textual information, and thus, should be recorded, at least on a documentation level.

« [...] media play their part in shaping utterances from the very beginning, they not only determine which signs we use but they also have an influence on how we use them. In short : Media offer a frame that, in the process of utterance production already, has an influence on how we design the utterance, how we process signs » (Luginbühl, 2015)

Guernsey Greffe, *Crime I*, f. 1



Anglo-Norman Year Books



- 1. CL & NLP
- 2. Challenges
- 3. Workflow
 - 3.1. Acquisition
 - 3.2. Structuring
 - 3.3. DataViz
- 4. Conclusion

3.2 Data structuring

- At least two levels, for CL/NLP to work : the *token* (word) level, and the *sentence* level.

```
<div n="1" type="section">
  <head>Prise de avers</head>
  <p n="1">
    <s n="1">
      <w n="1">[Johan]</w>
      <w n="2">de</w>
      <w n="3">Walton</w>
      <w n="4">porta</w>
      <w n="5">une</w>
      <w n="6">prise</w>
      <w n="7">de</w>
      <w n="8">avers</w>
      <w n="9">vers</w>
      <w n="10">W.</w>
      <w n="11">de</w>
      <w n="12">Seynt</w>
      <w n="13">Martyn</w>
      <w n="14">.</w>
    </s>
  </p>
```

Python/ReGex


```
5 #div n="1" type="section">
6 #head>Prise de avers</head>
7 #p n="1">
8 #s n="1">
9 1 [Johan]
10 2 de
11 3 Walton
12 4 porta
13 5 une
14 6 prise
15 7 de
16 8 avers
17 9 vers
18 10 W.
19 11 de
20 12 Seynt
21 13 Martyn
22 14 .
```

Project description

CoNLL-U Parser

CoNLL-U Parser parses a [CoNLL-U formatted](#) string into a nested python dictionary. CoNLL-U is often the output of natural language processing tasks.

Why should you use conllu?

- It's simple. ~300 lines of code.
- It has no dependencies
- Full typing support so your editor can do autocompletion
- Nice set of tests with CI setup:  passing
- It has 100% test branch coverage (and has undergone [mutation testing](#))
- It has [downloads 1M](#)

<https://pypi.org/project/conllu/>

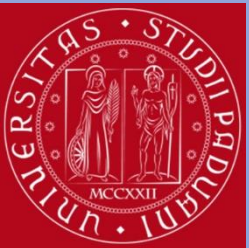
- You can add comments to the CONLLU (with lines beginning with #), and it is important for future synchronization (based on token numbering). As such, the XML-TEI can be more precise than the CONLL/PSD.

```
#on ouvre le fichier d'origine, et on remplace les < par des  
  
reading_file = open(path_in, "r")  
  
new_file_content = ""  
for line in reading_file:  
    stripped_line = line.strip()  
    diese = stripped_line.replace("<", "#")  
    new_file_content += diese + "\n"  
reading_file.close()  
  
writing_file = open(path_out, "w")  
writing_file.write(new_file_content)  
writing_file.close()
```

```
30 dire.  
#/s>  
#foreign xml:lang="la">  
#hi rend="italic">Initium sancti Euangelii sectndant Joh  
principio erat Verbum</hi>; </foreign>  
#s n="4">  
1 &  
2 comme  
3 nous  
4 continuyons  
5 à  
6 dire  
7 lad.  
8 éuangille,  
9 le  
10 corps  
11 de  
12 lad.  
13 Française  
14 qui
```

```
import xml.etree.ElementTree as ET  
  
coord={}  
roots=[]  
  
with open(xmlfile) as xmlfile:  
  
    # On importe le XML-TEI d'entrée et on le lit.  
    tree = ET.parse(xmlfile)  
    root = tree.getroot()  
  
    # On cible les chapter, et on récupère l'attribut n  
  
    for chapter in root.findall('.//div[@type="chapter"]'):  
        chapter_nb = chapter.get('n')  
  
        # On cible les sections, et on récupère l'attribut n  
  
        for section in chapter.findall('.//div[@type="section"]'):  
            section_nb = section.get('n')  
  
            #On cible les paragraphes, et on récupère l'attribut n  
  
            for para in section.findall('.//p'):  
                para_nb = para.get('n')  
  
                #On cible les sentences, et on récupère l'attribut n  
  
                for sentence in para.findall('.//s'):  
                    sentence_nb = sentence.get('n')
```

- CL & NLP
- Challenges
- Workflow
 - Acquisition
 - Structuring
 - Visualizat°
- Conclusion



3.2 Data structuring

- Mixed approach with NLP models, Treebanks, dictionary, etc.

```

1 [Johan] [johan] PROPN      _      _      4      nsubj      _      _
2 de un ADP      _      _      1      flat       _      PrestoPOS=Dn
3 Walton walton PROPN      _      _      1      flat       _      _
4 porta porter VERB       _      _      0      root        _      PrestoPOS=Vvc
5 une une DET       _      _      6      det         _      PrestoPOS=Nc
6 prise priser NOUN      _      _      4      obj         _      PrestoPOS=Vvc
7 de un ADP       _      _      8      case        _      PrestoPOS=Dn
8 avers avers NOUN      _      _      6      nmod        _      PrestoPOS=Nc
9 vers vert ADP       _      _      10     case        _      PrestoPOS=Nc
10 W. w. PROPN     _      _      4      obl         _      _
11 de un ADP       _      _      13     case        _      PrestoPOS=Dn
12 Seynt _ ADJ       _      _      13     amod        _      _
13 Martyn martyn PROPN    _      _      10     nmod        _      _
14 . . ADV       _      _      10     advmod     _      PrestoPOS=Fs
#/s>

```

=> Structuration = Standardization

```

<w lemma="porter" function="0:root" n="6" pos="VJ|Vvc">porta</w>
<ph function="IP-obj">
  <w lemma="un" function="8:det" n="7" pos="D|Dn">une</w>
  <w lemma="prise" function="6:obj" n="8" pos="NCS|Nc">prise</w>
</ph>
<ph function="IP">
  <w lemma="de" function="10:case" n="9" pos="P|S">de</w>
</ph>
  <w lemma="avers" function="8:nmod" n="10" pos="NCS|Nc">avers</w>
</ph>
</ph>
<ph function="IP">

```

- CL & NLP
- Challenges
- Workflow
- 3.1. Acquisition
- 3.2. Structuring**
- 3.3. Visualizat°
- Conclusion

- ID: Word index, integer starting at 1 for each new sentence; may be a range for multiword tokens; may be a decimal number for empty nodes (decimal numbers can be lower than 1 but must be greater than 0).
- FORM: Word form or punctuation symbol.
- LEMMA: Lemma or stem of word form.
- UPOS: [Universal part-of-speech tag](#).
- XPOS: Language-specific part-of-speech tag; underscore if not available.
- FEATS: List of morphological features from the [universal feature inventory](#) or from a defined [language-specific extension](#); underscore if not available.
- HEAD: Head of the current word, which is either a value of ID or zero (0).
- DEPREL: [Universal dependency relation](#) to the HEAD ([root](#) iff HEAD = 0) or a defined language-specific subtype of one.
- DEPS: Enhanced dependency graph in the form of a list of head-deprel pairs.
- MISC: Any other annotation.

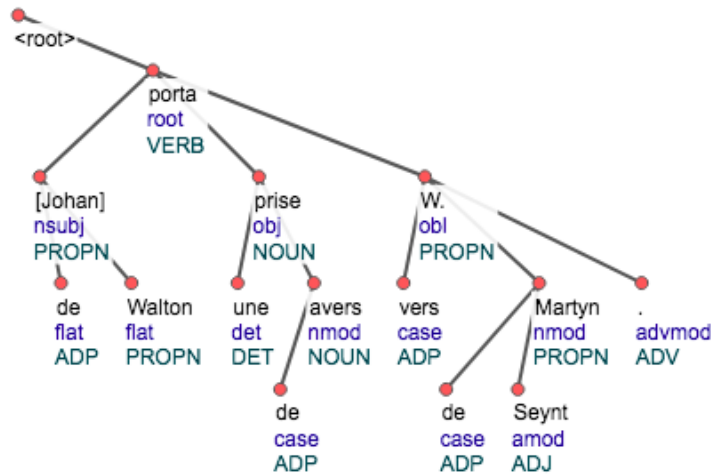
<https://universaldependencies.org/docs/format.html>

3.3 Data visualization

=> To be able to display *every encoding level*, from the word to the text.

From the conllu:

[Johan] de Walton porta une prise de avers vers W. de Seynt Martyn .



▼ C La seconde distinction, — Lire tout le chapitre.

C Du duc. xij. : « LE duc de normendie ou le prince est cil

C De aliance. xiiij. : « LE duc doit lauoir lalian ce et la lor

C De feaulte. xiiij. : « TOus ceulx qui sont res seantz eu di
... »

C De monneage. xv. De monneage etc. : « LE monneage e
deue au ... »

C De mesures. xvij. De mesures etc. : « Toute la poo ste et
mesures ... »

From the XML-TEI:

Requête : [pos="*.VJ.*"]

	Référence	Contexte gauche	Pivot	Contexte droit
122	7_1406_Gauvard	a la mort comme on dit le descoulpa.	Conclut	que il fait a recevoir et alias ut supra et que on
123	7_1406_Gauvard	Regnault sera examiné et au surplus la court	verra	tout ce que les parties vaudront en conseil et en arrest. A@
124	7_1406_Gauvard	la court verra tout ce que les parties	vaudront	en conseil et en arrest.
125	7_Murano_custom	@le ghiesie de questa terra de Muran,	siano	ben custodidi et non vadino in sinistro, però sia statuido et
126	7_Murano_custom	de Muran, siano ben custodidi et non	vadino	in sinistro, però sia statuido et ordenado che per cadauna ghiesia
127	7_Murano_custom	de@ @la ghiesia. Li qual do ellecti	siano	per miser lo piovàn et sacerdoti predicti significadi a miser lo pode
128	7_Murano_custom	a miser lo podestà, a@ @l qual	stia	la confirmation over revocation de tal ellectione, né altramente far

1. CL & NLP
2. Challenges
3. Workflow
- 3.1. Acquisition
- 3.2. Structuring
- 3.3. DataViz
4. Conclusion

4. References & closing statements

Our objectives, in short:

- To combine philological & linguistic approaches.
- To enable data analysis at least at three levels (word/sentence/text)
- To offer as many search options (as much as possible!)
- To stabilise annotation protocols and workflows suitable for all natural languages

1. CL & NLP
2. Challenges
3. Workflow
4. Conclusion

Camps, J-B et al. (2020), « Corpus and Models for Lemmatisation and POS-tagging of Classical French Theatre. 2020. halshs-02591388 (Preprint)

Gabay, S. et al. (2020), « CORPUS17: a philological corpus for 17th c. French », *DDH '20, 15-17/10 2020*, Hammamet

Gabay, S. et al. (2020), « Standardizing linguistic data: method and tools for annotating (pre-orthographic) French », , *DDH '20, 15-17/10 2020*, Hammamet

Ide, N. (2008), « Preparation and Analysis of Linguistic Corpora », *in* Schreibman (2008)

Luginbühl, M. (2015), « Media Linguistics: On Mediality and Culturality », *Media Linguistics*, 1.

Manning, D. (2015), « Computational Linguistics and Deep Learning », *in* *Computational Linguistics*, 41/4

Schreibman, S. et al. (2008), *A Companion to Digital Humanities*. Blackwell.

