



# Annotation collaborative de corpus : Formats

Karën Fort

karen.fort@sorbonne-universite.fr / <https://members.loria.fr/KFort/>



## Formats d'annotation

Annotation en ligne vs déportée

Format : késako ?

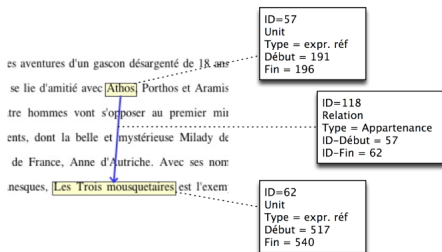
Des formats à connaître

Des formats aux schémas

Pour finir

# Où est l'annotation ?

- ▶ dans le fichier lui-même : annotation *inline*  
(*'The', 'AT'*), (*'Fulton', 'NP-TL'*), (*'County', 'NN-TL'*),  
(*'Grand', 'JJ-TL'*), (*'Jury', 'NN-TL'*), (*'said', 'VBD'*) [[Brown corpus](#)]
- ▶ dans un fichier séparé : annotation déportée (*standoff*)



[[Glozz](#)]

Quel(s) formats d'annotation connaissez-vous ?



# Formats linéaires

('The', 'AT'), ('Fulton', 'NP-TL'), ('County', 'NN-TL'), ('Grand', 'JJ-TL'), ('Jury', 'NN-TL'), ('said', 'VBD') [Brown corpus]

The DT the  
TreeTagger NP TreeTagger  
is VBZ be  
easy JJ easy  
to TO to  
use VB use  
. SENT .

PAT : <boy [\*] no>[/]/] girl [/] girl truck # girl +... [CHILDES]

⇒ simple, mais peu d'expressivité (interprétation nécessaire)

# TEI (Text Encoding Initiative)

Vous connaissez !  
(voir cours de M1)

# eXtensible Markup Language (XML)

en 20 sec.

- ▶ langage informatique de balisage (comme HTML ou SGML)
- ▶ ... textuel, structuré, et extensible car
- ▶ son « langage » (vocabulaire et grammaire) peut être redéfini (par exemple, *mabalise* peut être un nom de balise)
- ▶ syntaxe stricte, peut être validée par des outils automatiques

## XML : extrait de TCOF-POS

```
<?xml version="1.0" encoding="UTF-8"?>
<document>
<loc nb="L2">
<w lemme="L2" pos="LOC">L2</w>
<w lemme="ben" pos="INT">ben</w>
<w lemme="on" pos="PRO:cls">on</w>
<w lemme="attendre" pos="VER:pres">attend</w>
<w lemme="on" pos="PRO:cls">on</w>
<w lemme="attendre" pos="VER:pres">attend</w>
<w lemme="qui" pos="PRO:int">qui</w>
<w lemme="normalement" pos="ADV">normalement</w>
</loc>
```



# La *Text Encoding Initiative* (TEI)

en 20 sec. (voir <http://www.tei-c.org>)

Consortium à but non lucratif :

- ▶ auto-financé
- ▶ constitué d'institutions, de projets de recherche et de chercheurs (64 membres)
- ▶ qui existe depuis 1987
- ▶ qui développe et maintient un **standard pour la représentation des textes numériques** : un format SGML au début, XML maintenant
- ▶ outre la documentation du format, la TEI fournit des outils et des formations

# La TEI : exemple

Wikipédia, TEI, Le Cid

## Acte II, Scène 2

**DON RODRIGUE** À moi, Comte, deux mots.

**LE COMTE** Parle.

**DON RODRIGUE** Ôte-moi d'un doute.

Connais-tu bien Don Diègue ?

**LE COMTE** Oui.

**DON RODRIGUE** Parlons bas, écoute.

Sais-tu que ce vieillard fut la même vertu,

La vaillance et l'honneur de son temps ? Le sais-tu ?

# La TEI : exemple

## Wikipédia, TEI

```
<div type="Act" n="I"><head>Acte II</head>
  <div type="Scene" n="1"><head>Scène 2</head>
    <sp><speaker>Rodrigue</speaker>
      <l part="i">À moi, comte, deux mots.</l></sp>
    <sp><speaker>Comte</speaker>
      <l part="m">Parle</l></sp>
    <sp><speaker>Rodrique</speaker>
      <l part="f">Ôte-moi d'un doute</l></sp>
    <sp><speaker>Comte</speaker>
      <l part="i">Connais-tu bien Don Diège ?</l></sp>
    <sp><speaker>Comte</speaker>
      <l part="m">Oui</l></sp>
    <sp><speaker>Rodrigue</speaker>
      <l part="f">Parlons bas, écoute.</l>
      <l>Sais-tu que ce vieillard fut la même vertu,</l>
      <l>La vaillance et l'honneur de son temps ? Le sais-tu ?</l></sp>
    ...
  </div>
  ...
</div>
```

## TEI (Text Encoding Initiative) : caractéristiques

- + distingue entre les pratiques obligatoires, les pratiques recommandées et les pratiques optionnelles
- + ? permet aux utilisateurs d'étendre les schémas de base

# (X)CES : Corpus Encoding Standard

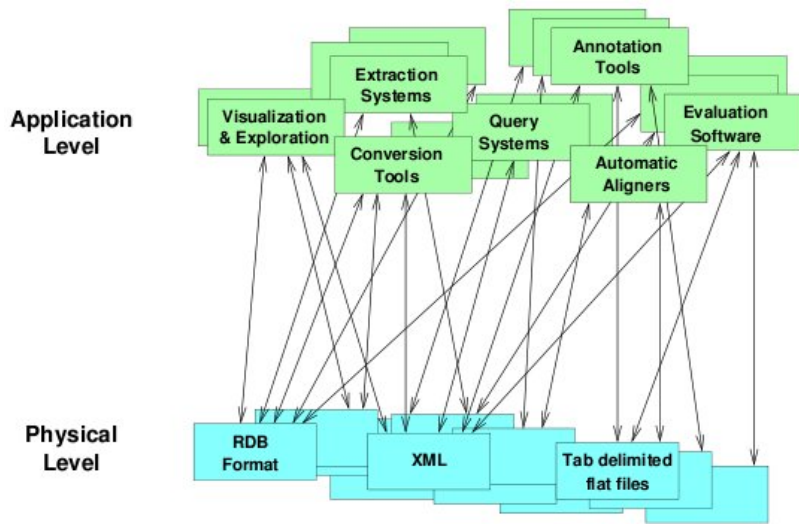
- + étend la TEI pour fournir un seul format de représentation pour les annotations **linguistiques** :
  - catégories génériques comme <msd> (*morpho-syntactic description*), avec la catégorie dans l'attribut ou le contenu de l'étiquette!
  - ⇒ les spécifications concernant la description des catégories linguistiques sont laissées aux bons soins de projets comme EAGLES/ISLE (duquel CES faisait partie)

++ annotation **déportée**

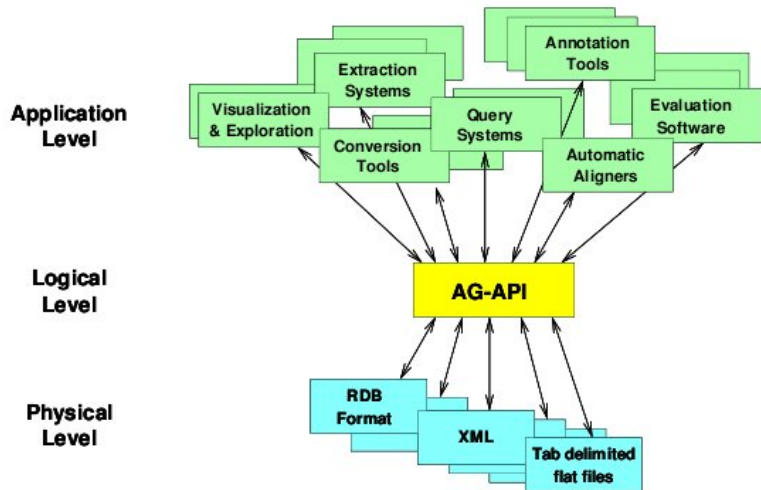
## [Bird and Liberman, 2001]

- ▶ les formats de **fichiers**, les **étiquettes** et les **attributs** sont secondaires
- ▶ la **structure logique** des annotations prime
- inspiré des BD :
  - ▶ interopérabilité
  - ▶ création et manipulation d'annotations selon votre tâche/besoin/préférences
  - ▶ principe d'indépendance des données

## D'architectures à 2 niveaux...



... à une architecture à 3 niveaux

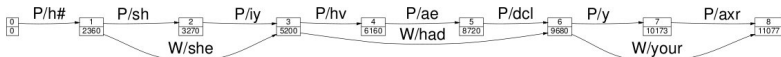




# Graphes d'annotation pour TIMIT

```
train/dr1/fjsp0/sal.wrd:  
2360 5200 she  
5200 9680 had  
9680 11077 your  
11077 16626 dark  
16626 22179 suit  
22179 24400 in  
24400 30161 greasy  
30161 36150 wash  
36720 41839 water  
41839 44680 all  
44680 49066 year
```

```
train/dr1/fjsp0/sal.phn:  
0 2360 h#  
2360 3720 sh  
3720 5200 iy  
5200 6160 hv  
6160 8720 ae  
8720 9680 dcl  
9680 10173 y  
10173 11077 axr  
11077 12019 dcl  
12019 12257 d  
...
```



# Graphes d'annotation pour UTF

```
<turn speaker="Roger_Hedgecock" spkrtype="male" dialect="native"
  startTime="2348.811875" endTime="2391.606000" mode="spontaneous" fidelity="high">
  ...
  <time sec="2378.629937">
  now all of those things are in doubt after forty years of democratic rule in
  <b_enamex type="ORGANIZATION">congress</b_enamex>
  <time sec="2382.539437">
  {breath because <contraction e_form="[you=>you] ['ve=>have]">you've got quotas
  {breath and set<hyphen>asides and rigidities in this system that keep you
  <time sec="2387.353875">
  on welfare and away from real ownership
  {breath and <contraction e_form="[that=>that] ['s=>is]">that's a real problem in this
  <b_overlap startTime="2391.115375" endTime="2391.606000">country</b_overlap>
</turn>
<turn speaker="Gloria_Allred" spkrtype="female" dialect="native"
  startTime="2391.299625" endTime="2439.820312" mode="spontaneous" fidelity="high">
  <b_overlap startTime="2391.299625" endTime="2391.606000">well i</b_overlap>
  think the real problem is that %uh these kinds of republican attacks
  <time sec="2395.462500">
  i see as code words for discrimination
  ...
</turn>
```



# Graphes d'annotation pour la Coréférence

<COREF ID="2" MIN="woman">This woman</COREF> receives three hundred dollars a month under <COREF ID="5">General Relief</COREF>, plus <COREF ID="16" MIN="four hundred dollars"> four hundred dollars a month in <COREF ID="17" MIN="benefits" REF="16">A.F.D.C. benefits</COREF></COREF> for <COREF ID="9" MIN="son"><COREF ID="3" REF="2">her</COREF>son</COREF>, who is <COREF ID="10" MIN="citizen" REF="9">a U.S. citizen</COREF>.

<COREF ID="4" REF="2">She</COREF>'s among <COREF ID="18" MIN="aliens">an estimated five hundred illegal aliens on <COREF ID="6" REF="5">General Relief</COREF> out of <COREF ID="11" MIN="population"><COREF ID="13" MIN="state">the state</COREF>'s total illegal immigrant population of <COREF ID="12" REF="11"> one hundred thousand </COREF></COREF></COREF> <COREF ID="7" REF="5">General Relief</COREF> is for needy families and unemployable adults who don't qualify for other public assistance. Welfare Department spokeswoman Michael Reganburg says <COREF ID="15" MIN="state" REF="13">the state</COREF> will save about one million dollars a year if <COREF ID="20" MIN="aliens" REF="18">illegal aliens</COREF> are denied <COREF ID="8" REF="5">General Relief</COREF>.



# Graphes d'annotation (Annotation Graphs)

[Bird and Liberman, 2001]

- ▶ graphes orientés acycliques  $\Rightarrow$  expressivité
- ▶ avec des enregistrements sur les arcs
- ▶ avec des références temporelles optionnelles sur les nœuds

# Linguistic Annotation Framework [Ide and Romary, 2006]

- ▶ norme ISO
- ▶ vise à :
  1. s'adapter à **tous les types** d'annotations linguistiques
  2. fournir les moyens de représenter des informations linguistiques **complexes**

# Principes de LAF

- ▶ séparation entre les **données** (*read-only*) et les **annotations** (*stand-off*)
  - ▶ séparation entre **le format d'annotation de l'utilisateur** et **le format d'échange**
  - ▶ séparation entre **la structure** et **le contenu** dans le format d'échange (liste = liste d'alternatives ou avec inclusion ou avec priorités?)
- ⇒ annotation = graphe orienté, instancié en XML (TEI)

## GrAF : application de LAF

Alors que les graphes d'annotation (AG) permettent de représenter des **couches** d'annotation, chacune associée à des données primaires...

... GrAF permet des annotations **reliées** à d'autres annotations (des annotations multiple forment un seul graphe)

Formats d'annotation

## Des formats aux schémas

Rappel sur les arbres

Syntaxe ou sémantique ?

Outils et formats

Pour finir



# Formats vs schémas

TEI

est du XML

# Formats vs schémas

TEI

est du XML

structure d'arbre ?

# Formats vs schémas

TEI

est du XML

structure d'arbre ?

LAF

est un graphe orienté acyclique

# Formats vs schémas

TEI

est du XML

structure d'arbre ?

LAF

est un graphe orienté acyclique

structure de graphe ?

# Formats vs schémas

TEI

est du XML

structure d'arbre ?

LAF

est un graphe orienté acyclique

structure de graphe ?

en TEI ??

XML : syntaxe ou sémantique ?



# Définition

## Arbre (théorie des graphes)

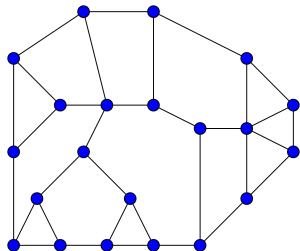
« un arbre est un graphe non orienté, acyclique et connexe »  
[Wikipédia, Arbre\_(graphe), consultée le 12/11/14]

# Définition

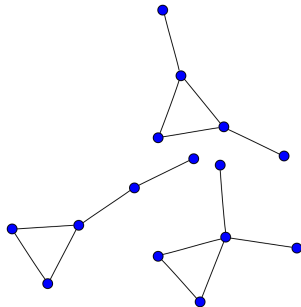
## Arbre (théorie des graphes)

« un arbre est un graphe non orienté, acyclique et connexe »  
[Wikipédia, Arbre\_(graphe), consultée le 12/11/14]

Graphe connexe (et cyclique) :



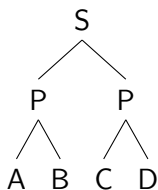
Graphe non connexe :



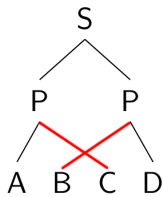
[Koko90]



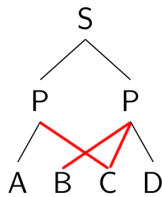
# Arbre vs graphe



Arbre  
(XML)



Arbre (non adjacence)  
(XML décorrélé)



Graphe (cyclique)  
(XML décorrélé)

# Structure vs Interprétation

- ▶ XML permet de représenter à la fois des arbres et des graphes
- ▶ l'interprétation est dans la structure ou en dehors de la structure : **expressivité**
- ▶ l'expressivité de XML est limitée :  $\Rightarrow$  utilisation d'annotations **stand-off** (décorrelées)

## Que choisit-on ?

Est-ce qu'on choisit un **format** d'annotation ?

## Que choisit-on ?

Est-ce qu'on choisit un **format** d'annotation ?

ou

un **outil** d'annotation (sans trop se préoccuper de la question du format) ?

## Question

GATE

Quel format utilise GATE?

## Question

### GATE

Quel format utilise GATE ?

*Note that GATE's model of annotation allows graph structures, which are difficult to represent in XML (XML is a tree-structured representation format). During the dump process, annotations that cross each other in ways that cannot be represented in legal XML will be discarded, and a warning message printed.*

<https://gate.ac.uk/sale/tao/splitch3.html#sec:developer:dump>

Formats d'annotation

Des formats aux schémas

**Pour finir**

CQFR : Ce Qu'il Faut Retenir  
TD



- ▶ des arbres aux graphes : évolution vers plus d'expressivité
- ▶ reste de la place pour plus de méthodologie !





Bird, S. and Liberman, M. (2001).

A formal framework for linguistic annotation.

Speech Communication, 33(1-2) :23–60.



Ide, N. and Romary, L. (2006).

Representing linguistic corpora and their annotations.

In Proceedings of the International Conference on Language Resources and Evaluation (LREC), Gène, Italie.