CC BY NC SA

Manual Annotation in NLP: manual annotation formats and tools

Karën Fort

karen.fort@univ-lorraine.fr / https://members.loria.fr/KFort

IDMC Institut des sciences du Digital Management & Cognition
COMPOSANTE DE L'UNIVERSITÉ DE LORRAINE

UNIVERSITÉ DE LORRAINE

Annotation formats
Inline vs standoff Annotation
Format: What are we talking about?
Some formats to know about

From formats to schemes

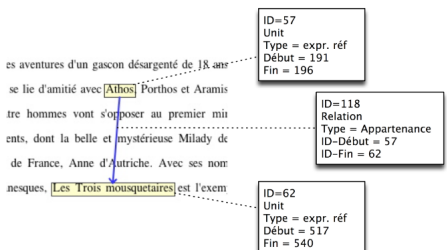
Annotation tools

Pre-annotation

To finish

Where are the annotations?

- in the file itself: *inline* annotation
  ('The', 'AT'), ('Fulton', 'NP-TL'), ('County', 'NN-TL'), ('Grand', 'JJ-TL'), ('Jury', 'NN-TL'), ('said', 'VBD') [Brown corpus]
- in a separate file: standoff annotation



[Glozz]

Which (manual) annotation formats do you know?

Linear formats

('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]

$\Rightarrow$ simple, but little expressivity (necessary interpretation)

eXtensible Markup Language (XML)

in 20 sec.

- markup language (like HTML)
- . . . textual, structured, and extensible as
- its "language" (vocabulary and grammar) can be redefined (eg, *mytag* can be the name of a tag)
- strict syntax that can be validated by automatic tools

XML: extract from TCOF-POS (French speech annotated with 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>
```

The *Text Encoding Initiative* (TEI)

in 20 sec. (see http://www.tei-c.org)

Non profit consortium:

- self-financed
- formed of institutions, research projectand researchers
- existing since 1987
- develops and maintains a standard for the representation of digitaliz(ed) texts: at first in SGML, now XML
- besides the format documentation, TEI provides tools and training

The TEI: example

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 ?

The TEI: example

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 Diègue ?</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>
...
```

TEI (Text Encoding Initiative) : characteristics

+ distinguishes between mandatory practices, recommended practices and optional practices

+? allows the users to extend the basic schemes

(X)CES: Corpus Encoding Standard

\+ extends TEI to provide one representation format for **linguistic** annotations:

- generci categories like <msd> (*morpho-syntactic description*), with the category in the features or in the tag content

⇒ specifications concerning the description of linguistic categories are taken care of by projects like EAGLES/ISLE (CES was part of it)

++ standoff annotation

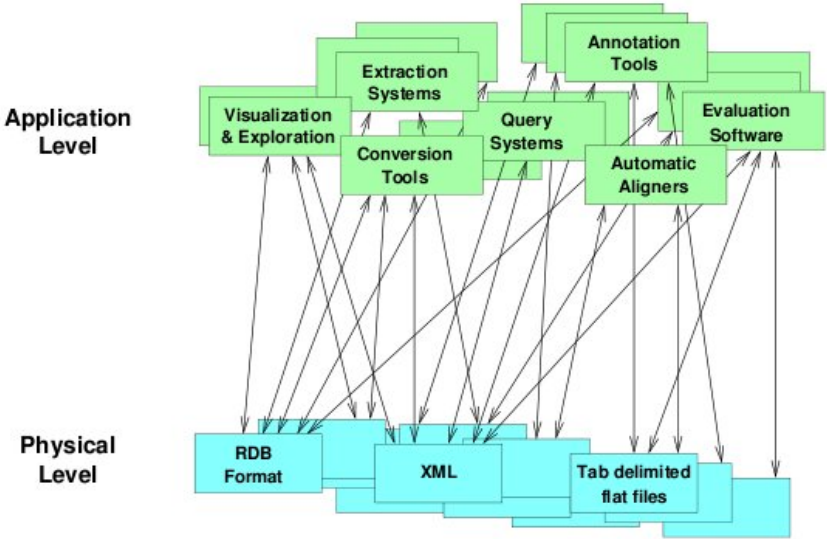
[Bird and Liberman, 2001]

- file formats, tags/labels and features are second
- the logical structure of the annotations comes first

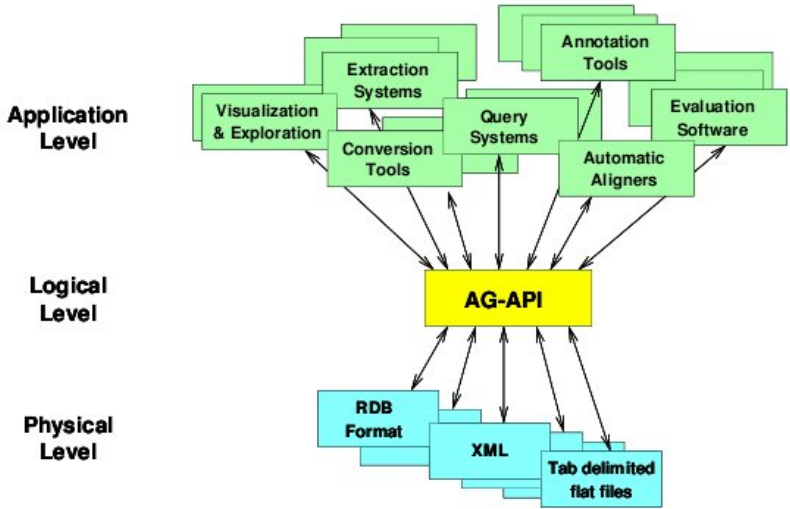
→ inspired from DB:

- interoperability
- creation and manipulation of annotations according to your task/need/preference
- principle of data independence

From a 2-level architecture...



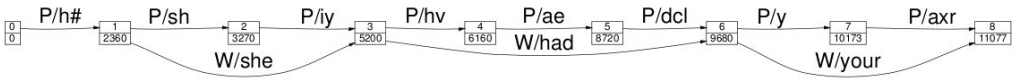
... to a 3-level architecture



Annotation graphs for TIMIT

```
train/dr1/fjsp0/sa1.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/sa1.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
...
```

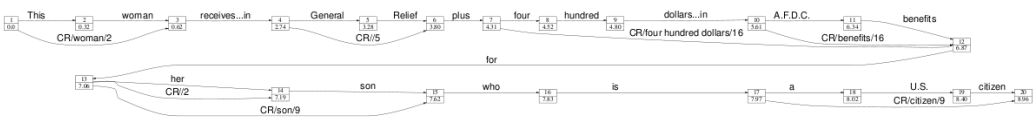


Annotation graphs for UTF

```
<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<e_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<e_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<e_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>
```

Annotation graphs for coreference

```
<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>.
```



Annotation graphs [Bird and Liberman, 2001]

- direct acyclic graphs $\Rightarrow$ expressivity
- with recordings on the arcs
- with optional temporal references on the nodes

Linguistic Annotation Framework [Ide and Romary, 2006]

- ISO norm
- aims at:
  1. being adaptable to all types of linguistic annotations
  2. providing the means to represent complex linguistic information

LAF principles

- separation between the data (*read-only*) and the annotations (*stand-off*)
- separation between the user's annotation format and the exchange format
- separation between the structure and the content in the exchange format (list = list of alternatives or with inclusion or with priorities?)

$\Rightarrow$ annotation = direct graph, instanciated in XML (TEI)

GrAF: application of LAF

While annotation graphs (AG) allow to represent several layers of annotations, each being associated to primary data...
... GrAF allows for annotations to be linked to other annotations (multiple annotations forming a unique graph)

Annotation formats

From formats to schemes
- Reminder about trees
- Syntax or semantics?
- Tools and formats

Annotation tools

Pre-annotation

To finish

Formats *vs* schemes

TEI
is in XML

Formats *vs* schemes

TEI
is in XML
tree structure?

Formats *vs* schemes

TEI
is in XML
tree structure?

LAF
is a direct acyclic graph

Formats *vs* schemes

TEI
is in XML
tree structure?

LAF
is a direct acyclic graph
graph structure?

Formats *vs* schemes

TEI
is in XML
tree structure?

LAF
is a direct acyclic graph
graph structure?
in TEI??

XML: syntax or semantics?

Definition

Tree (graph theory)

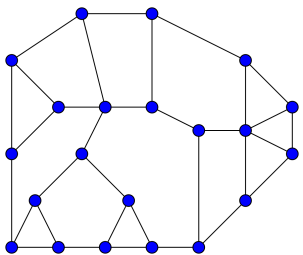
"In graph theory, a tree is an undirected graph in which every pair of distinct vertices is connected by exactly one path, or equivalently, a connected acyclic undirected graph" [Wikipedia, Tree (graph theory), consulted on Sept. 1st, 2025]

Definition

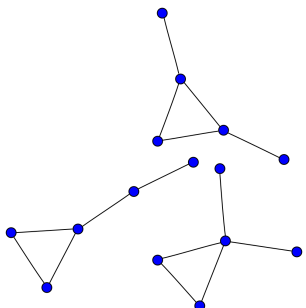
Tree (graph theory)

"In graph theory, a tree is an undirected graph in which every pair of distinct vertices is connected by exactly one path, or equivalently, a connected acyclic undirected graph" [Wikipedia, Tree (graph theory), consulted on Sept. 1st, 2025]

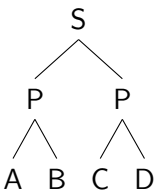
connected (and cyclic) graph:



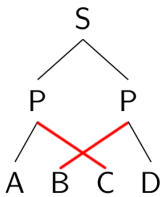
unconnected graph:



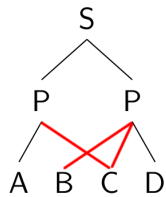
Tree *vs* graph



Tree
(XML)



Tree (non adjacent)
(decorelated XML)



Graph (cyclic)
(decorelated XML)

Structure *vs* Interpretation

- XML allows to represent **both** trees and graphs
- the interpretation is in the structure or outside of the structure: expressivity
- the expressivity of XML is limited: $\Rightarrow$ need to use stand-off annotations

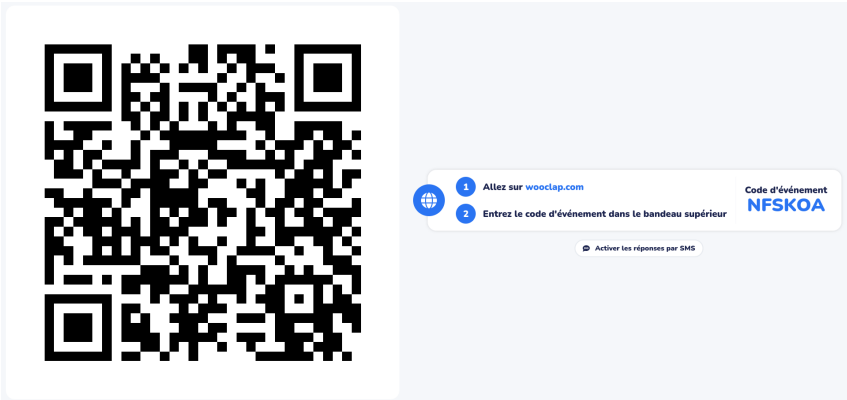
How do we choose?

Do we choose an annotation format?

How do we choose?

Do we choose an annotation format?
or
an annotation tool (without caring so much about the format)?

Wooclap: What is the most used format today in NLP?



CoNLL-U Format

```
1   They     they    PRON   PRP   Case=Nom|Number=Plur              2   nsubj   2:nsubj|4:nsubj
2   buy      buy     VERB   VBP   Number=Plur|Person=3|Tense=Pres   0   root    0:root
3   and      and     CCONJ  CC    _                                 4   cc      4:cc
4   sell     sell    VERB   VBP   Number=Plur|Person=3|Tense=Pres   2   conj    0:root|2:conj
5   books    book    NOUN   NNS   Number=Plur                       2   obj     2:obj|4:obj
6   .        .       PUNCT  .     _                                 2   punct   2:punct
```

CoNLL-U

Annotation formats

From formats to schemes

Annotation tools

Pre-annotation

To finish

Solutions to the manual annotation cost?

Solutions to the manual annotation cost?

- Annotation tools
- Tag dictionnaries / Pre-annotation / active learning
- Training / Documentation / Methodologie
- crowdsourcing: `Amazon Mechanical Turk` and games with a purpose (`GWAPs`)

What for?

Why using tools?

- to ease the edition of annotations, in particular in the case of relations

Why using tools?

- to ease the edition of annotations, in particular in the case of relations
- to limit the number of items keep in mind [Dandapat et al., 2009]

Why using tools?

- to ease the edition of annotations, in particular in the case of relations
- to limit the number of items keep in mind [Dandapat et al., 2009]
- to constrain the annotation, and thus limit the errors
  [de la Clergerie, 2008, Mikulová and Štěpánek, 2009]

Why using tools?

- to ease the edition of annotations, in particular in the case of relations
- to limit the number of items keep in mind [Dandapat et al., 2009]
- to constrain the annotation, and thus limit the errors
  [de la Clergerie, 2008, Mikulová and Štěpánek, 2009]
- to hide a layer when annotating another [Widlöcher and Mathet, 2009]

Why using tools?

- to ease the edition of annotations, in particular in the case of relations
- to limit the number of items keep in mind [Dandapat et al., 2009]
- to constrain the annotation, and thus limit the errors
  [de la Clergerie, 2008, Mikulová and Štěpánek, 2009]
- to hide a layer when annotating another [Widlöcher and Mathet, 2009]
- to ease access to the context, even large [Widlöcher and Mathet, 2009]

Why using tools?

- to ease the edition of annotations, in particular in the case of relations
- to limit the number of items keep in mind [Dandapat et al., 2009]
- to constrain the annotation, and thus limit the errors
  [de la Clergerie, 2008, Mikulová and Štěpánek, 2009]
- to hide a layer when annotating another [Widlöcher and Mathet, 2009]
- to ease access to the context, even large [Widlöcher and Mathet, 2009]
- to keep track of the discussions between annotators [Lortal et al., 2006] or of the errors and their correction [de la Clergerie, 2008]

Some existing tools

+/- WebAnno, Glozz, GATE, Knowtator, Callisto, etc.

++ gain in time and quality

$\Rightarrow$ (too) many tools, developed to demonstrate the interest of an annotation scheme or for a specific annotation campaign, not for the annotators

Annotation formats

From formats to schemes

Annotation tools

Pre-annotation

To finish

Tag Dictionaries

Allow to:

1. store the categories associated to a token by the annotators
2. propose these categories when the same token is to be annotated again

$\Rightarrow$ Very simple and relatively efficient (see [Carmen et al., 2010]), but the more you annotate, the more the method is efficient

Correction of automatic pre-annotations

++ significant gain in time and quality, at least for POS tagging and syntactic annotation (`Penn Treebank` [Marcus et al., 1993], POS tagging for Hindi and Bangla [Dandapat et al., 2009], POS tagging for English [Fort and Sagot, 2010])

- bias not always taken into account: is it the same to pre-annotate named entities and gene renaming relations?
- also time consuming if the system is not performant enough

Sub-category of pre-annotation: Active Learning

- all the annotations are not necessary to train a tool⇒ detect the annotations which are really useful to improve the final results
- automatically pre-annotare a corpus, then ask the annotators to correct the annotation, then re-train the tool and identify, using the scores, what needs to be modify, etc.

⇒ iterative

\+ allows to gain time

\- but time consuming if the system is not performant enough

- on the Rital project (oral human-machine dialog): more than 30% of errors, the transcribers worked faster from scratch

Questions about pre-annotation

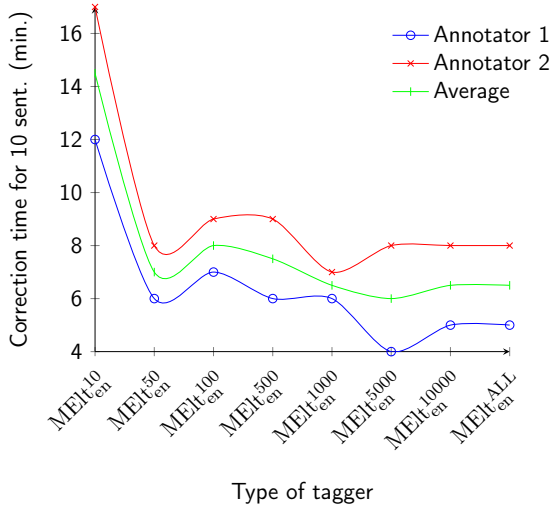
- either the humans focus on what has been pre-annotated and correct the pre-annotations, without seeing what's missing
- or they focus on what's missing and they don't correct the pre-annotation.
- impossible for some annotation types because of the lack of high quality systems (like co-reference chains annotators)

Impact of the pre-annotation [Fort and Sagot, 2010]

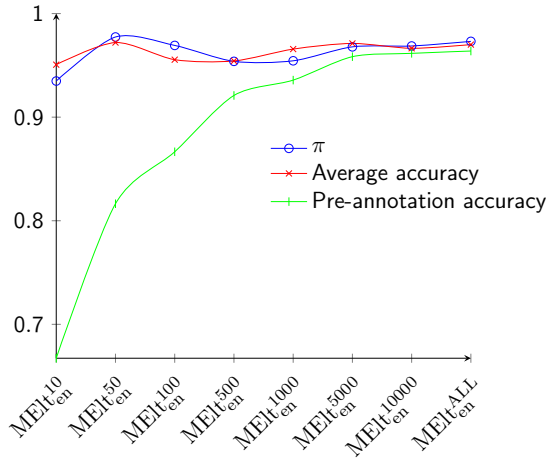
- gain in time and quality (inter-annotator agreement and accuracy)
- influence of different levels of quality of the pre-annotation
- bias introduced by the pre-annotation

  ... while limiting the **effects of the learning curve**

Correction time

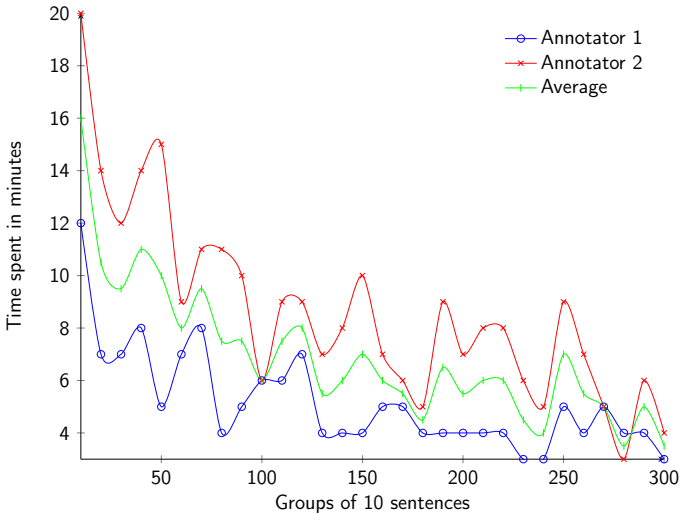


Correction quality



Size of the training corpus for the tagger

Learning curve: POS annotation of the *Penn Treebank* [Fort and Sagot, 2010]



Training and documentation

A good training of the annotators is the most efficient solution to gain annotation time and quality [Dandapat et al., 2009].

This has to be associated with an adapted documentation proposing:

- a clear definition of the **application**
- a clear and detailed definition of the **categories** (always possible or even a good idea?)
- some well-chosen **examples**
- a separate presentation of the **ambiguous categories**, like in the PTB documentation (see: ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz)

Training and documentation

A good training of the annotators is the most efficient solution to gain annotation time and quality [Dandapat et al., 2009].

This has to be associated with an adapted documentation proposing:

- a clear definition of the **application**
- a clear and detailed definition of the **categories** (always possible or even a good idea?)
- some well-chosen **examples**
- a separate presentation of the **ambiguous categories**, like in the PTB documentation (see: ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz)

Do not forget that the annotators are at the heart of the annotation!

Annotation formats

From formats to schemes

Annotation tools

Pre-annotation

To finish
WYMR: What You Must Remember
Practice

- Formats
- Tools
- Impact of pre-annotation

Practice: Using Inception

https://inception.grew.fr/login.html

- create one project per team (M1_2026_X, with X being the initials of the team members)
- add the other users/team members to the project (ids are M1_2026_NAME, passwords are 4NLPMaster2026)
- add Karën Fort and Lucie Digoin-Caparros to the project
- set up the tool to:
  - annotate your corpus with POS
  - plan the overlaps between annotators
  - compute the inter-annotator agreement

Bird, S. and Liberman, M. (2001).
A formal framework for linguistic annotation.
Speech Communication, 33(1-2):23–60.

Carmen, M., Felt, P., Haertel, R., Lonsdale, D., McClanahan, P., Merkling, O., Ringger, E., and Seppi, K. (2010).
Tag dictionaries accelerate manual annotation.
In Proceedings of the International Conference on Language Resources and Evaluation (LREC), La Valette, Malte. European Language Resources Association (ELRA).

Dandapat, S., Biswas, P., Choudhury, M., and Bali, K. (2009).
Complex linguistic annotation - no easy way out! a case from bangla and hindi POS labeling tasks.
In Proceedings of the third ACL Linguistic Annotation Workshop, Singapour.

de la Clergerie, E. V. (2008).
A collaborative infrastructure for handling syntactic annotations.

In Proceedings of the First International Workshop on Automated Syntactic Annotations for interoperable Language Resources, Hong-Kong, Chine.

Fort, K. and Sagot, B. (2010).
Influence of pre-annotation on POS-tagged corpus development.
In Proceedings of the Fourth ACL Linguistic Annotation Workshop, pages 56–63, Uppsala, Suède.

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.

Lortal, G., Todirascu-Courtier, A., and Lewkowicz, M. (2006).
Soutenir la coopération par l'indexation semi-automatique d'annotations.
In Actes de la Semaine de la Connaissance 2006, Nantes, France.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993).
Building a large annotated corpus of English : The Penn Treebank.
Computational Linguistics, 19(2):313–330.

Mikulová, M. and Štěpánek, J. (2009).
Annotation quality checking and its implications for Design of treebank (in building the prague czech-english Dependency treebank).
In Proceedings of the Eight International Workshop on Treebanks and Linguistic Theories, volume 4-5, Milan, Italie.

Widlöcher, A. and Mathet, Y. (2009).
La plate-forme Glozz : environnement d'annotation et d'exploration de corpus.
In Actes de Traitement Automatique des Langues Naturelles (TALN), Senlis, France.