Manual Annotation:
What is it? How to do it (properly)?

Karën Fort

karen.fort@loria.fr / https://members.loria.fr/KFort/

École thématique d’été "Annotations" - May 31st, 2022
Qual Program

https://members.loria.fr/KFort/publications/tutorials-summer-schools-etc/

- **Qual1**: now
- **Qual2** (Thursday, 11 am): inter-annotator agreement
- **Qual3** (Friday, 9 am): crowdsourcing
Where I’m talking from
https://members.loria.fr/KFort/

- Manual annotation for NLP

- Ethics and NLP
Manuel Annotation and NLP

What is Annotation?

How to do this properly?

Analysing the complexity of an annotation campaign

To finish
Manuel Annotation and NLP
  Manuel Annotation in NLP
  A notoriously high cost
  About language resources longevity

What is Annotation?

How to do this properly?

Analysing the complexity of an annotation campaign

To finish
NLP today

Massive Language data

Language model

Manual annotation

Annotated corpus

Reference corpus

Evaluation

Annotated corpus

Application

Used by

Language data

Model

Used by

Language data

Produce

Used by

Produce

Used by

Produce
NLP today

Massive Language data

Language model produce

Manual annotation

Annotated corpus

Reference corpus

Evaluation

Annotated corpus

Application

Language data

Model

Used by

Produce

Used by
Manual annotation in NLP
Manual annotation in for NLP
Why it’s important

Ben Hamner @benhamner · Oct 9

Programming: 10% writing code. 90% figuring out why it doesn’t work

Analyzing data and ML: 1% writing code. 9% figuring out why code doesn’t work. 90% figuring out what’s wrong with the data
PTB 1:

- POS-tagging correction: ? units per hour, ? hours a day
- constituents (syntax) correction: ? units per hour, ? hours a day
PTB 1:

- POS-tagging correction: 3,000 units per hour, ? hours a day
- constituents (syntax) correction: ? units per hour, ? hours a day
Penn Treebank (PTB) [Marcus et al., 1993]

PTB 1:
- POS-tagging correction: 3,000 units per hour, 3 hours a day
- constituents (syntax) correction: ? units per hour, ? hours a day
Penn Treebank (PTB) [Marcus et al., 1993]

PTB 1:
- POS-tagging correction: 3,000 units per hour, 3 hours a day
- constituents (syntax) correction: 750 units per hour, ? hours a day
Penn Treebank (PTB) [Marcus et al., 1993]

PTB 1:

- POS-tagging correction: 3,000 units per hour, 3 hours a day
- constituents (syntax) correction: 750 units per hour, 3 hours a day
Penn Treebank (PTB) [Marcus et al., 1993]

PTB 1:

- POS-tagging correction: 3,000 units per hour, 3 hours a day
- constituents (syntax) correction: 750 units per hour, 3 hours a day
- + learning curve 1 month (POS-tagging) to 2 months (syntax)!
1996-2004 [Böhmová et al., 2001],
built from the CNC (Czech National Corpus),
3 structural levels:
1. morphological (semi-automatic): 1.8 million tokens
2. analytical (dependency syntax, with an ad-hoc tool)
3. tectogrammatical (semantic): 1 million tokens
Prague Dependency Treebank [Böhmová et al., 2001]

Version 1.0:
- manual annotation of the morphological and analytical levels
- time: ?
- nb of persons: ?
- cost estimate: ?
Prague Dependency Treebank [Böhmová et al., 2001]

Version 1.0:
- manual annotation of the morphological and analytical levels
- time: 5 years
- nb of persons: ?
- cost estimate: ?
Prague Dependency Treebank [Böhmová et al., 2001]

Version 1.0:
- manual annotation of the morphological and analytical levels
- time: 5 years
- nb of persons: 22 persons, incl. 17 simultaneously during the most demanding periods
- cost estimate: ?
Prague Dependency Treebank [Böhmová et al., 2001]

Version 1.0:

- manual annotation of the morphological and analytical levels
- time: 5 years
- nb of persons: 22 persons, incl. 17 simultaneously during the most demanding periods
- cost estimate: $600,000
GENIA [Kim et al., 2008]

**GENIA**: 400,000 words annotated in microbiology.
GENIA [Kim et al., 2008]

**GENIA**: 400,000 words annotated in microbiology.

⇒ 5 part-time annotators, 1 senior coordinator, 1 junior coordinator during 1.5 year [Kim et al., 2008]
100 h of transcribed speech (ESTER evaluation campaign, transcription systems, 2008)
1 h of speech = ?
100 h of transcribed speech (ESTER evaluation campaign, transcription systems, 2008)

1 h of speech = between 20 and 60 h of transcription work
100 h of transcribed speech (ESTER evaluation campaign, transcription systems, 2008)

1 h of speech = between 20 and 60 h of transcription work

⇒ quality has to be high!
Lifespan of annotated corpora

*Penn Treebank* [Marcus et al., 1993]:
- created at the beg. of the 90s
- still used in 2022 (ACL 2022)

vs PARTS tagger [Church, 1988], which is no more used

→ fast evolution of the tools

⇒ annotation should **not** depend on them
What is Annotation?
  Practice
  DefinitionS
  What for?

How to do this properly?

Analysing the complexity of an annotation campaign

To finish
Practice

Transcribe what you hear, using any text editor or paper.
“[corpus annotation] can be defined as the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data. ’Annotation’ can also refer to the end-product of this process” [Leech, 1997]

“’Linguistic annotation’ covers any descriptive or analytic notations applied to raw language data. The basic data may be in the form of time functions - audio, video and/or physiological recordings - or it may be textual.” [Bird and Liberman, 2001]
Adding interpretative information [Leech, 1997, Habert, 2005]

Annotation

ANNOTATION

SEGMENTATION

Signal
Annotation

Adding interpretative information [Leech, 1997, Habert, 2005]
The application: horizon of the annotation

An annotation is always task-oriented [Habert, 2000].

- direct applicative purpose (summaries of football matches for the football campaign)
- intermediary application or internal to NLP application (POS-tagging)

The annotations are more useful, the more they are designed to be specific to a particular application [Leech, 2005].
With a huge surprise from the side of Bayern Munich as Van Bommel, the captain, has been removed. He is not even on the substitutes list.
Exercice: annotate soccer match comments
players, teams, actions (goals), relations (passes), etc.

With a huge surprise from the side of Bayern Munich as Van Bommel, the captain, has been removed. He is not even on the substitutes list.

What is the task, the application aimed at?

summary of match

Van Bommel?

should not be annotated
The consensus, at the heart of annotation

One needs to "agree to be able to measure" [Desrosières, 2008]

Annotation is related to quantification

Measuring vs quantifying [Desrosières, 2008] :

- **measuring**: implies a measurable form (eg. the height of Mont Blanc)
- **quantifying**: implies preliminary conventions of equivalence

The consensus should be equipped:

- annotation guidelines (12p. for soccer)
- meetings with the annotators and the campaign manager
- evaluate the consensus (consistency)
Manuel Annotation and NLP

What is Annotation?

How to do this properly?
   Good Practises
   Theorizing

Analysing the complexity of an annotation campaign

To finish
Leech’s 7 maxims [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Note: can be hard after normalization (“l’arbre” → “le arbre”, etc.)
Leech’s 7 maxims [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Note: can be hard after normalization (“l’arbre” → “le arbre”, etc.)

2. Annotations should be extractable from the text
Leech’s 7 maxims [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Note: can be hard after normalization (“l’arbre” → “le arbre”, etc.)
2. Annotations should be extractable from the text
3. The annotation procedure should be documented (ex: Brown Corpus annotation guide, Penn Tree Bank annotation guide)
Leech’s 7 maxims [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Note: can be hard after normalization (“l’arbre” → “le arbre”, etc.)

2. Annotations should be extractable from the text

3. The annotation procedure should be documented (ex: Brown Corpus annotation guide, Penn Tree Bank annotation guide)

4. Mention should be made of the annotator(s) and the way annotation was made (manual/automatic annotation, number of annotators, manually corrected/uncorrected...)


Leech’s 7 maxims [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Note: can be hard after normalization (“l’arbre” → “le arbre”, etc.)
2. Annotations should be extractable from the text
3. The annotation procedure should be documented (ex: Brown Corpus annotation guide, Penn Tree Bank annotation guide)
4. Mention should be made of the annotator(s) and the way annotation was made (manual/automatic annotation, number of annotators, manually corrected/uncorrected…)
5. Annotation is an act of interpretation (cannot be infallible)
Leech’s 7 maxims [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Note: can be hard after normalization (“l’arbre” \(\rightarrow\) “le arbre”, etc.)
2. Annotations should be extractable from the text
3. The annotation procedure should be documented (ex: Brown Corpus annotation guide, Penn Tree Bank annotation guide)
4. Mention should be made of the annotator(s) and the way annotation was made (manual/automatic annotation, number of annotators, manually corrected/uncorrected...)
5. Annotation is an act of interpretation (cannot be infallible)
6. Annotation schemas should be as independent as possible on formalisms
Leech’s 7 maxims [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Note: can be hard after normalization ("l’arbre" → "le arbre", etc.)

2. Annotations should be extractable from the text

3. The annotation procedure should be documented (ex: Brown Corpus annotation guide, Penn Tree Bank annotation guide)

4. Mention should be made of the annotator(s) and the way annotation was made (manual/automatic annotation, number of annotators, manually corrected/uncorrected...)

5. Annotation is an act of interpretation (cannot be infallible)

6. Annotation schemas should be as independent as possible on formalisms

7. No annotation schema should consider itself a standard (it possibly becomes one)
Different points of view

“you only get out what you put in” [Wallis, 2007]
Model-based approach

Knowledge is in the annotation schema $\Rightarrow$ corpus comes after

Everything is in the annotation!
Corpus-based approach

Knowledge is in the text $\Rightarrow$ the corpus comes first [Sinclair]
Third way?

The knowledge is in the annotation schema and in the corpus.
Annotation by cycles

- new observations generalize hypotheses
- theory allows to interpret and classify information
- evolving cycles: each cycle improves the knowledge by refining and testing the theories on real data
  ⇒ a more precise representation of the corpus is built and a more sophisticated system is produced
Agile Annotation

integrating evaluation

Traditionnal annotation phases (left) and cycles of agile annotation (right). Reproduction of Figure 2 from [Voormann and Gut, 2008]
Generic annotation pipeline [Hovy and Lavid, 2010]
MATTER cycle [Pustejovsky and Stubbs, 2012]

1. Model
2. Annotate
3. Train
4. Test
5. Evaluate
6. Revise
Towards "annotation engineering" [Fort, 2012]
Manuel Annotation and NLP

What is Annotation?

How to do this properly?

Analysing the complexity of an annotation campaign
  What to annotate?
  How to annotate?
  Weight of the context
  Synthesis

To finish
Complexity dimensions

- 5 independent dimensions:
  - 2 related to the localisation of annotations
  - 3 related to the characterisation of annotations
  - 1 not independent: the context

- Scale from 0 (null complexity) to 1 (maximal complexity) to allow for the comparison between campaigns

- Independent from the volume to annotate and the number of annotators
Example: gene renaming

1. Identification of gene names in the source signal:
   The *yppB* gene complemented the defect of the recG40 strain. *yppB* and *ypbC* and their respective null alleles were termed “*recU*” and “*recU1*” (*recU*:cat) and “*recS*” and “*recS1*” (*recS*:cat), respectively.

2. Identification of gene couples expressing a renaming relation:
   The *yppB* gene complemented the defect of the recG40 strain. *yppB* and *ypbC* and their respective null alleles were termed “*recU*” and “*recU1*” (*recU*:cat) and “*recS*” and “*recS1*” (*recS*:cat), respectively.
Discrimination

Parts-of-speech [Marcus et al., 1993], pre-annotated:

*I/PRP do/VBP n’t/RB feel/VB very/RB ferocious/JJ ./.

Gene renaming [Fort et al., 2012], no pre-annotation:

The *yppB:cat* and *ypbC:cat* null alleles rendered cells sensitive to DNA-damaging agents, impaired plasmid transformation (25- and 100-fold), and moderately affected chromosomal transformation when present in an otherwise Rec+ *B. subtilis* strain. The *yppB* gene complemented the defect of the *recG40* strain. *yppB* and *ypbC* and their respective null alleles were termed *recU* and “*recU1*” (*recU:cat*) and *recS* and “*recS1*” (*recS:cat*), respectively. The *recU* and *recS* mutations were introduced into rec-deficient strains representative of the alpha (*recF*), beta (*addA5 addB72*), gamma (*recH342*), and epsilon (*recG40*) epistatic groups.
Discrimination

Parts-of-speech [Marcus et al., 1993], pre-annotated:

*I PRP do/VBP n’t/RB feel/VB very/RB ferocious/JJ ./.

Gene renaming [Fort et al., 2012], no pre-annotation:

The yppB:cat and ypbC:cat null alleles rendered cells sensitive to DNA-damaging agents, impaired plasmid transformation (25- and 100-fold), and moderately affected chromosomal transformation when present in an otherwise Rec+ B. subtilis strain. The yppB gene complemented the defect of the recG40 strain. yppB and ypbC and their respective null alleles were termed recU and “recU1” (recU:cat) and recS and “recS1” (recS:cat), respectively. The recU and recS mutations were introduced into rec-deficient strains representative of the alpha (recF), beta (addA5 addB72), gamma (recH342), and epsilon (recG40) epistatic groups.

⇒ more difficult if the units to annotate are scattered, in particular if the segmentation is not obvious.
Discrimination

The discrimination weight is all the more high as the proportion of what \textit{should} be annotated as compared to what \textit{could} be annotated is low.

**Definition**

\[
\text{Discrimination(Flow)} = 1 - \frac{\left| \text{Annotations(Flow)} \right|}{\sum_{i=1}^{\text{LevelSeg}} \left| \text{UnitsObtainedBySeg}_i(\text{Flow}) \right|}
\]

⇒ Need for a reference segmentation
Parts-of-speech [Marcus et al., 1993]:

\[ I/PRP \ do/VBP \ n’t/RB \ feel/VB \ very/RB \ ferocious/JJ ./ . \]

\[ Discrimination_{PTB_{POS}} = 0 \]

Gene renaming [Fort et al., 2012]:

The \textit{yppB:cat} and \textit{ypbC:cat} null alleles rendered cells sensitive to DNA-damaging agents, impaired plasmid transformation (25- and 100-fold), and moderately affected chromosomal transformation when present in an otherwise Rec+ \textit{B. subtilis} strain. The \textit{yppB} gene complemented the defect of the recG40 strain. \textit{yppB} and \textit{ypbC} and their respective null alleles were termed \textit{recU} and “\textit{recU1}” (\textit{recU:cat}) and \textit{recS} and “\textit{recS1}” (\textit{recS:cat}), respectively. The \textit{recU} and \textit{recS} mutations were introduced into rec-deficient strains representative of the alpha (\textit{recF}), beta (\textit{addA5 addB72}), gamma (\textit{recH342}), and epsilon (\textit{recG40}) epistatic groups.

\[ Discrimination_{Identification} = 0.9 \]
\[ Discrimination_{Renaming} = 0.95 \]
Boundaries delimitation

- **extending** or **shrinking** the discriminated unit:
  
  *Madame Chirac → Monsieur et Madame Chirac*
Boundaries delimitation

- extending or shrinking the discriminated unit:
  *Madame Chirac → Monsieur et Madame Chirac*

- decompose a discriminated unit into several elements:
  *le préfet Érignac → le *préfet* Érignac*
Boundaries delimitation

- **extending** or **shrinking** the discriminated unit: 
  *Madame Chirac* → *Monsieur et Madame Chirac*

- **decompose** a discriminated unit into several elements: 
  *le préfet Érignac* → *le préfet Érignac*

- **or group** together several discriminated units into one unique annotation:
  *Sa Majesté* 
  *le roi Mohamed VI* → *Sa Majesté le roi Mohamed VI*
Boundaries delimitation

Definition

\[ \text{Delimitation}(\text{Flow}) = \min \left( \frac{\text{Substitutions} + \text{Additions} + \text{Deletions}}{|\text{Annotations}(\text{Flow})|}, 1 \right) \]

\[ \text{Delimitation}_{\text{Identification}} = 0 \]
\[ \text{Delimitation}_{\text{Renaming}} = 0 \]

\[ \text{Delimitation}_{\text{PTBPOS}} = 0 \]

\[ \text{Délimitation}_{\text{ENTypesSubtypes}} = 1 \]
\[ \text{Délimitation}_{\text{ENComponents}} = 0, 3 \]
Expressiveness of the annotation language

Definition

The degrees of expressiveness of the annotation language are the following:

- 0.25: type languages
- 0.5: relational languages of arity 2
- 0.75: relational languages of arity higher than 2
- 1: higher-order languages

Expressiveness\textsubscript{Identification} = 0.25

Expressiveness\textsubscript{Renaming} = 0.25
### Dimension of the tagset

<table>
<thead>
<tr>
<th>Person</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{pers.ind}</td>
<td>\textit{func.ind}</td>
</tr>
<tr>
<td>(individual person)</td>
<td>(individual function)</td>
</tr>
<tr>
<td>\textit{pers.coll}</td>
<td>\textit{func.coll}</td>
</tr>
<tr>
<td>(group of persons)</td>
<td>(collectivity of functions)</td>
</tr>
</tbody>
</table>

#### Level 1:
- \textit{pers.ind}:
  - Administrative
    - (\textit{loc.adm.town}, \textit{loc.adm.reg}, \textit{loc.adm.nat}, \textit{loc.adm.sup})
  - Physical
    - (\textit{loc.phys.geo}, \textit{loc.phys.hydro}, \textit{loc.phys.astro})
  - Facilities
    - (\textit{loc.fac}), onynms
      - (\textit{loc.oro}), address
    - (\textit{loc.add.phys}, \textit{loc.add.elec})

#### Level 2:
- \textit{func.ind}:
  - Production
    - \textit{prod.object}
      - (manufac-tured object)
    - \textit{prod.serv}
      - (transporta-tion route)
    - \textit{prod.doctr}
      - (doctrine)
    - \textit{prod.rule}
      - (law)
    - \textit{prod.fin}
      - (financial products)
    - \textit{prod.soft}
      - (software)
    - \textit{prod.art}
      - (art)
    - \textit{prod.media}
      - (media)
    - \textit{prod.award}
      - (award)

#### Level 3:
- \textit{func.coll}:
  - Organization
    - (administration)
  - Time
    - \textit{time.date.abs}
      - (absolute date)
    - \textit{time.date.rel}
      - (relative date)

#### Types and sub-types used for structured NE annotation [Grouin et al., 2011]
## Dimension of the tagset

<table>
<thead>
<tr>
<th><strong>Person</strong></th>
<th><strong>Function</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>pers.ind (individual person)</td>
<td>func.ind (individual function)</td>
</tr>
<tr>
<td>pers.coll (group of persons)</td>
<td>func.coll (collectivity of functions)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Location</strong></th>
<th><strong>Production</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>administrative (loc.adm.town, loc.adm.reg, loc.adm.nat, loc.adm.sup)</td>
<td>physical (loc.phys.geo, loc.phys.hydro, loc.phys.astro)</td>
</tr>
<tr>
<td>facilities (loc.fac), oronyms (loc.oro), address (loc.add.phys, loc.add.elec)</td>
<td>prod.object (manufactured object)</td>
</tr>
<tr>
<td>prod.serv (transportation route)</td>
<td>prod.fin (financial products)</td>
</tr>
<tr>
<td>prod.doctr (doctrine)</td>
<td>prod.rule (law)</td>
</tr>
<tr>
<td>prod.soft (software)</td>
<td>prod.art</td>
</tr>
<tr>
<td>prod.media</td>
<td>prod.award</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Organization</strong></th>
<th><strong>Time</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>org.adm (administration)</td>
<td>time.date.abs (absolute date), time.date.rel (relative date)</td>
</tr>
<tr>
<td>org.ent (services)</td>
<td>time.hour.abs (absolute hour), time.hour.rel (relative hour)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Amount</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>amount (with unit or general object), including duration</td>
</tr>
</tbody>
</table>

**Level 1:** pers, func, loc, prod, org, time, amount $\rightarrow$ 7 possibilities (degree of freedom $= 6$).
Dimension of the tagset

<table>
<thead>
<tr>
<th>Person</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>pers.ind</td>
<td>func.ind</td>
</tr>
<tr>
<td>pers.coll</td>
<td>func.coll</td>
</tr>
<tr>
<td>person</td>
<td>function</td>
</tr>
</tbody>
</table>

| Level 1: pers, func, loc, prod, org, time, amount → 7 possibilities (degree of freedom = 6). |

<table>
<thead>
<tr>
<th>Location</th>
<th>Production</th>
</tr>
</thead>
<tbody>
<tr>
<td>administrative (loc.adm.town, loc.adm.reg, loc.adm.pat, loc.adm.sup)</td>
<td>physical (loc.phys.geo, loc.phys.hydro, loc.phys.astro)</td>
</tr>
<tr>
<td>facilities (loc.fac), oronym (loc.ofo), address (loc.add.phys, loc.add.elec)</td>
<td>prod.object (manufac-tured object)</td>
</tr>
<tr>
<td>prod.serv (transporta-tion route)</td>
<td>prod.fin (financial products)</td>
</tr>
<tr>
<td>prod.doctr (doctrine)</td>
<td>prod.rule (law)</td>
</tr>
<tr>
<td>prod.art</td>
<td>prod.media</td>
</tr>
<tr>
<td>prod.award</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Organization</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>org.adm (administra-tion)</td>
<td>org.ent (services)</td>
</tr>
<tr>
<td>time.date.abs (absolute date)</td>
<td>time.hour.abs (absolute hour), time.hour.rel (relative hour)</td>
</tr>
<tr>
<td>time.date.rel (relative date)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>amount (with unit or general object), including duration</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>time.date.abs (absolute date)</td>
</tr>
<tr>
<td>time.hour.abs (absolute hour)</td>
</tr>
<tr>
<td>time.hour.rel (relative hour)</td>
</tr>
</tbody>
</table>

Level 2: prod.object, prod.serv, prod.fin, prod.soft, prod.doctr, prod.rule, prod.art, prod.media, prod.award → 9 possibilities (degree of freedom = 8).
Dimension of the tagset

<table>
<thead>
<tr>
<th>Person</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>pers.ind</em> (individual person)</td>
<td><em>func.ind</em> (individual function)</td>
</tr>
<tr>
<td><em>pers.coll</em> (group of persons)</td>
<td><em>func.coll</em> (collectivity of functions)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Location</th>
<th>Production</th>
</tr>
</thead>
<tbody>
<tr>
<td>administrative (locadm.town, locadm.reg, locadm.nat, locadm.sup)</td>
<td><em>prod.object</em> (manufactured object)</td>
</tr>
<tr>
<td>physical (locphys.geo, locphys.hydro, locphys.astro)</td>
<td><em>prod.serv</em> (transportation route)</td>
</tr>
<tr>
<td>facilities (locfac), ononyms (locoro), address (locadd.phys, locadd.elec)</td>
<td><em>prod.fin</em> (financial products)</td>
</tr>
<tr>
<td>prod.doctr (doctrine)</td>
<td>prod.rule (law)</td>
</tr>
<tr>
<td>prod.art</td>
<td>prod.sofi (software)</td>
</tr>
<tr>
<td>prod.media</td>
<td>prod.award</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Organization</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>org.adm</em> (administration)</td>
<td><em>time.date.abs</em> (absolute date), <em>time.date.rel</em> (relative date)</td>
</tr>
<tr>
<td><em>org.ent</em> (services)</td>
<td><em>time.hour.abs</em> (absolute hour), <em>time.hour.rel</em> (relative hour)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>amount</em> (with unit or general object), including duration</td>
</tr>
</tbody>
</table>

Level 1: *pers, func, loc, prod, org, time, amount* $\rightarrow$ 7 possibilities (degree of freedom = 6).
Level 2: *prod.object, prod.serv, prod.fin, prod.soft, prod.doctr, prod.rule, prod.art, prod.media, prod.award* $\rightarrow$ 9 possibilities (degree of freedom = 8).
Level 3: *locadm.town, locadm.reg, locadm.nat, locadm.sup* $\rightarrow$ 4 possibilities (degree of freedom = 3).
Dimension of the tagset

Degree of freedom

\[ \nu = \nu_1 + \nu_2 + \ldots + \nu_m \]

where \( \nu_i \) is the maximal degree of freedom the annotator has when choosing the \( i^{th} \) sub-type (\( \nu_i = n_i - 1 \)).

Dimension of the tagset

Dimension(Flow) = \( \min(\frac{\nu}{\tau}, 1) \)

where \( \tau \) is the threshold from which we consider the tagset to be very large (experimentally determined).

\[ \begin{align*}
\text{Dimension}_{\text{Identification}} &= 0 \\
\text{Dimension}_{\text{Renaming}} &= 0.04 \\
\text{Dimension}_{\text{NETypesSubtypes}} &= 0.34
\end{align*} \]
Degree of ambiguity: residual ambiguity

Using the traces left by the annotators:

[...,] \(<\text{EukVirus}>3\text{CDproM}</\text{EukVirus}>\) can process both structural and nonstructural precursors of the \(<\text{EukVirus}\ \text{uncertainty-type} = "too-generic">\text{<taxon>poliovirus</taxon> polyprotein}</\text{EukVirus}>\) [...].

Définition

\[
\text{AmbiguityRes}(\text{Flow}) = \frac{|\text{Annotations}_{amb}|}{|\text{Annotations}|}
\]

\[
\text{AmbiguityRes}_{\text{Identification}} = 0.04
\]

\[
\text{AmbiguityRes}_{\text{Renaming}} = 0.02
\]
Degree of ambiguity: theoretical ambiguity

Proportion of the units to annotate that corresponds to ambiguous vocables.

Definition

\[
\text{AmbiguityTh}(Flow) = \sum_{voc_i=1}^{\mid \text{Voc}(Flow) \mid} \left( \text{Ambig}(voc_i) \ast \text{freq}(voc_i, Flow) \right) / \mid \text{Units}(Flow) \mid
\]

with

\[
\text{Ambig}(voc_i) = \begin{cases} 
1 & \text{if } \mid \text{Tags}(voc_i) \mid > 1 \\
0 & \text{else}
\end{cases}
\]

\[
\text{AmbiguityTh}_{\text{Identification}} = 0.01
\]

→ Does not apply to renaming relations
Context to take into account

▶ **size of the window** to take into account in the source signal:
  ▶ The sentence:
    
    \[ I/PRP \text{ do/VBP n't/RB feel/VB very/RB ferocious/JJ ./} \]
  ▶ ... or more:
    
    Fabien Lévêque : C'est bien fait, avec [Soulfran] maintenant. [Soulfran] qui va tenter sa chance, et ça fait le but. Le but !
    
    Xavier Gravelaine : Oh la la la la !
    

▶ **number of knowledge elements** to be rallied or degree of accessibility of the knowledge sources that are consulted:
  ▶ annotation guidelines
  ▶ nomenclatures (Swiss-Prot)
  ▶ new sources to be found (Wikipedia, etc.)
Weight of the context
Synthesis of the complexity dimensions

Classification of *it* pronouns as anaphoric or impersonal

Gene names identification
What’s next: Qual Program

https://members.loria.fr/KFort/publications/tutorials-summer-schools-etc/

- **Qual1**: now
- **Qual2** (Thursday, 11 am): inter-annotator agreement
- **Qual3** (Friday, 9 am): crowdsourcing
Manuel Annotation and NLP

What is Annotation?

How to do this properly?

Analysing the complexity of an annotation campaign

To finish

WYHTR: What You Have To Remember
Manual annotation and NLP:
▶ usage
▶ cost

Manual annotation:
▶ definition
▶ organization
▶ complexity grid
A bit of reading

The Manual Annotation Complexity Grid

Modeling the Complexity of Manual Annotation Tasks: a Grid of Analysis
Karën Fort, Adeline Nazarenko, Sophie Rosset
https://hal.archives-ouvertes.fr/hal-00769631/file/coling2012_Complexity_KF_30102012.pdf


Corpus annotation schemes.

Corpus annotation: Linguistic information from computer text corpora, chapter Introducing corpus annotation, pages 1–18.
Longman, Londres, Angleterre.


Building a large annotated corpus of English: The Penn Treebank.

Natural Language Annotation for Machine Learning.
O’Reilly.