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

LETTRES SORBONNE UNIVERSITÉ

gdr LIFT CNRS

Loria Laboratoire lorrain de recherche en informatique et ses applications

Qual Program

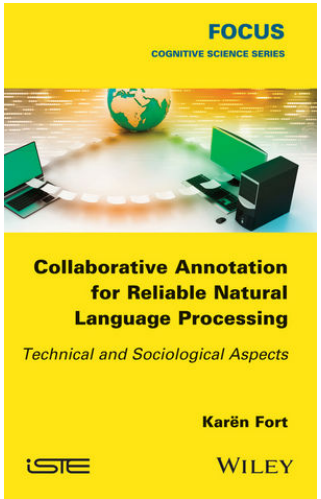
https://members.loria.fr/KFort/publications/tutorials-summer-schools-etc/

- 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

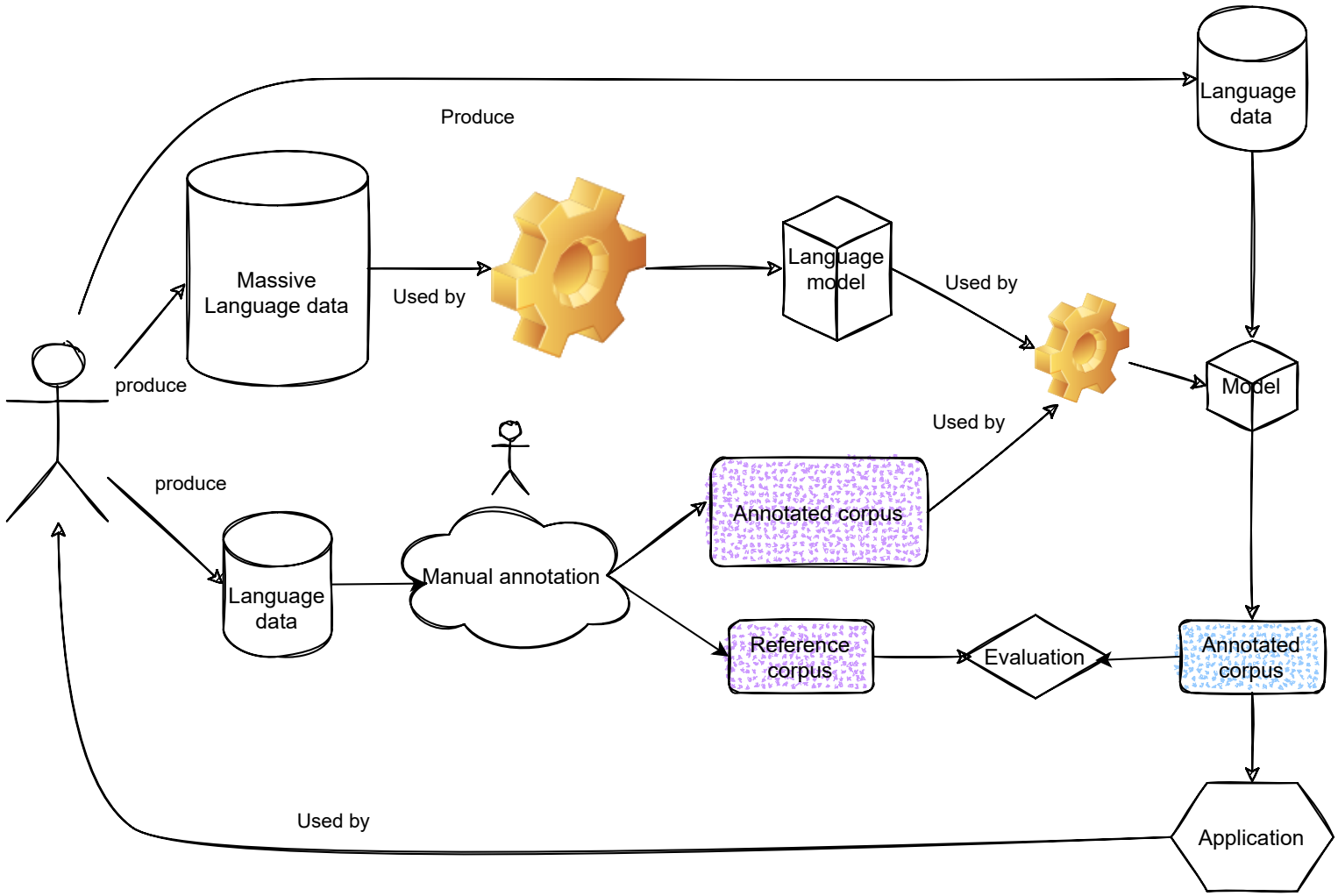
What is Annotation?

How to do this properly?

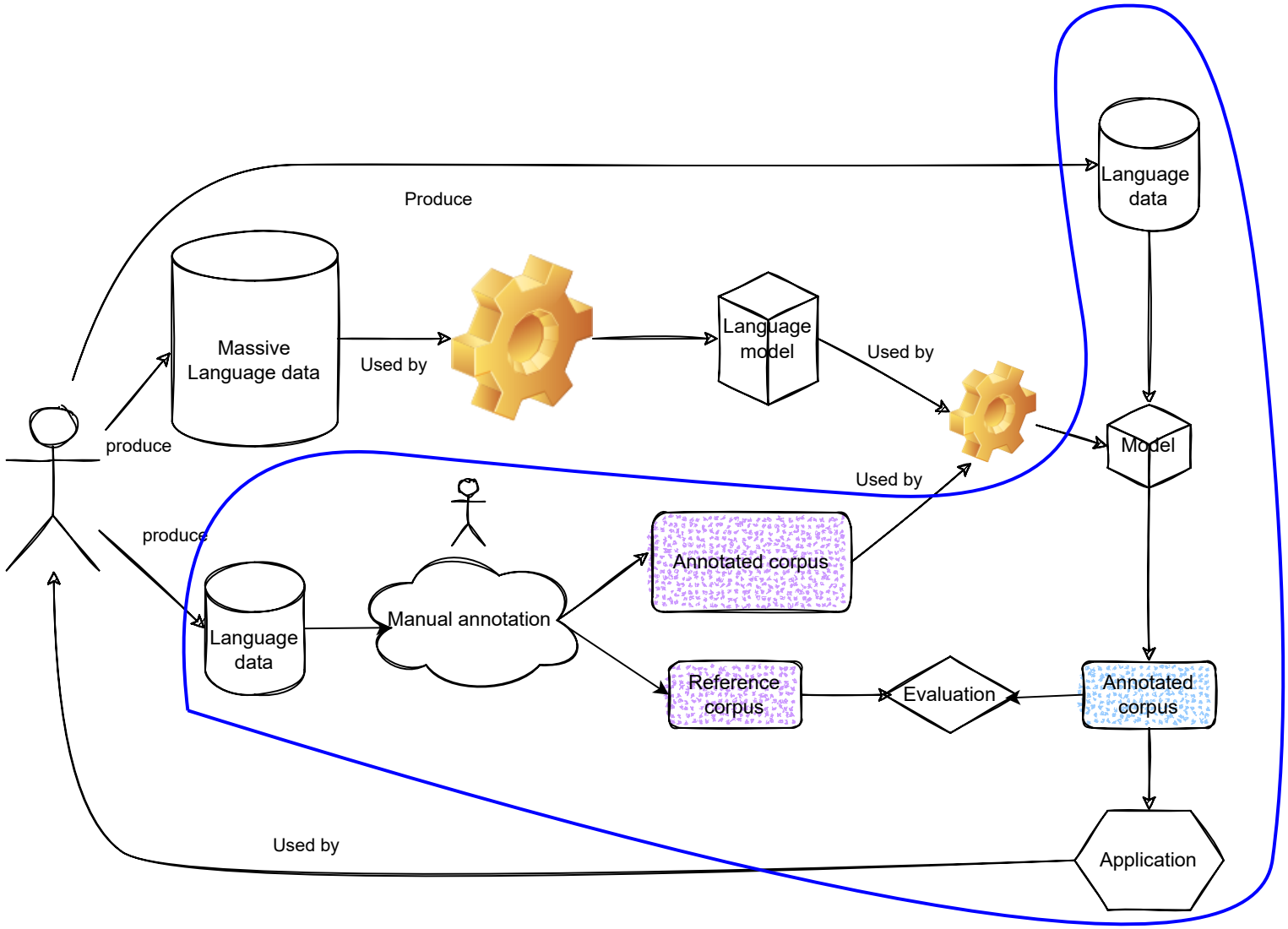
Analysing the complexity of an annotation campaign

To finish

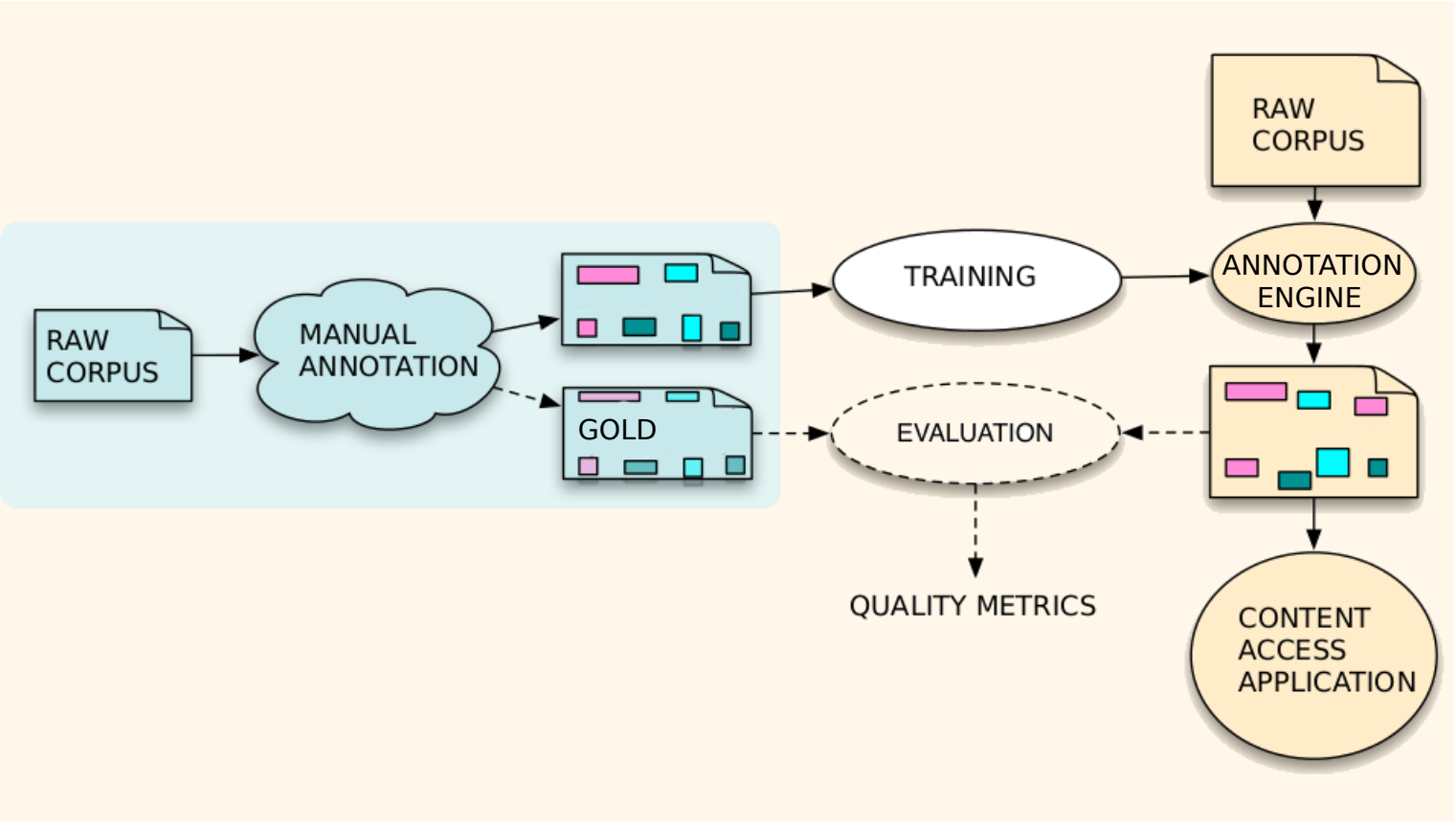
NLP today


Language data
Produce
Massive Language data
Used by
Language model
Used by
produce
Model
Used by
produce
Annotated corpus
Language data
Manual annotation
Reference corpus
Evaluation
Annotated corpus
Used by
Application

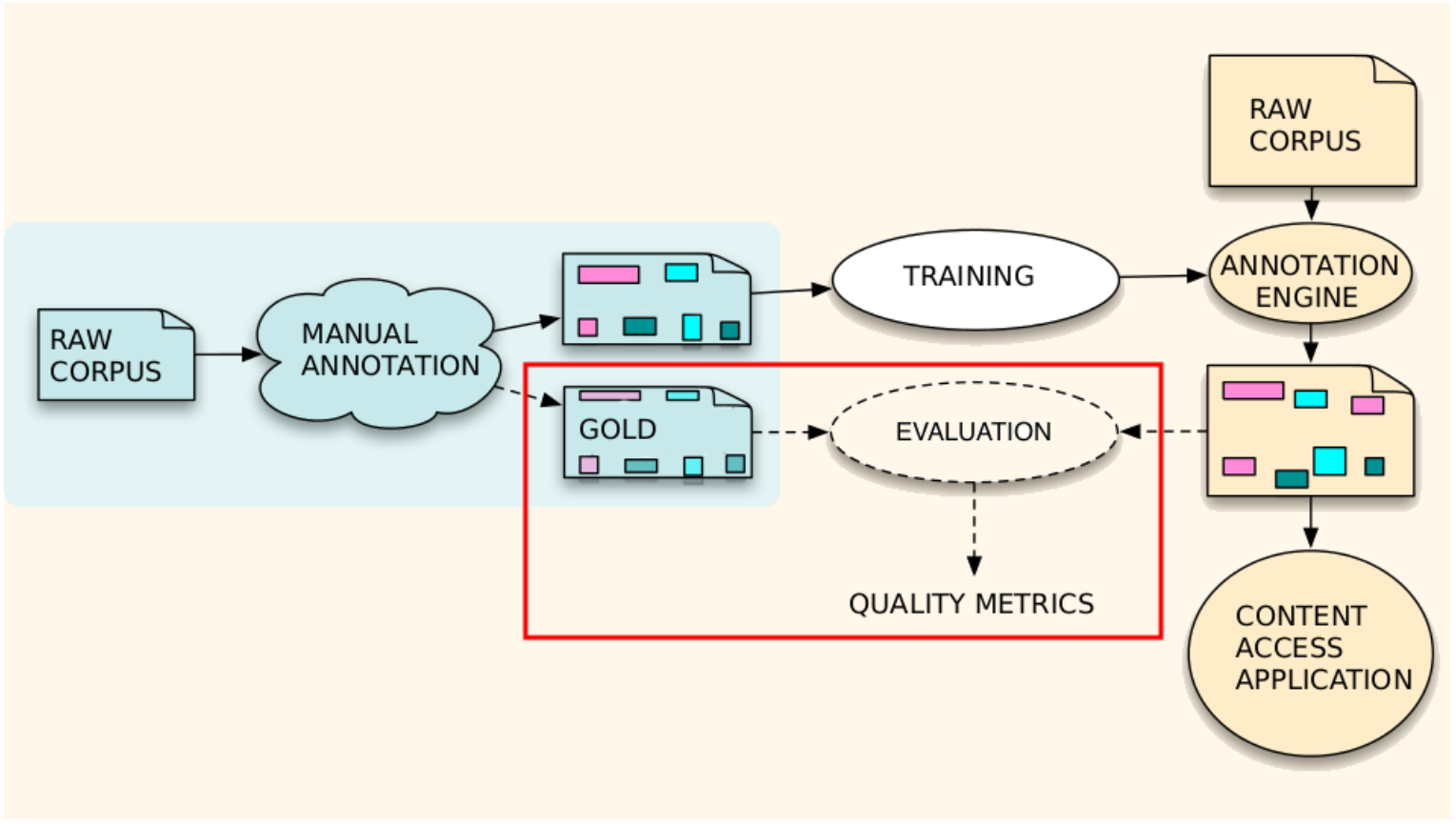
NLP today


Produce
Language data
Massive Language data
Used by
Language model
Used by
produce
Model
Used by
produce
Annotated corpus
Language data
Manual annotation
Reference corpus
Evaluation
Annotated corpus
Used by
Application

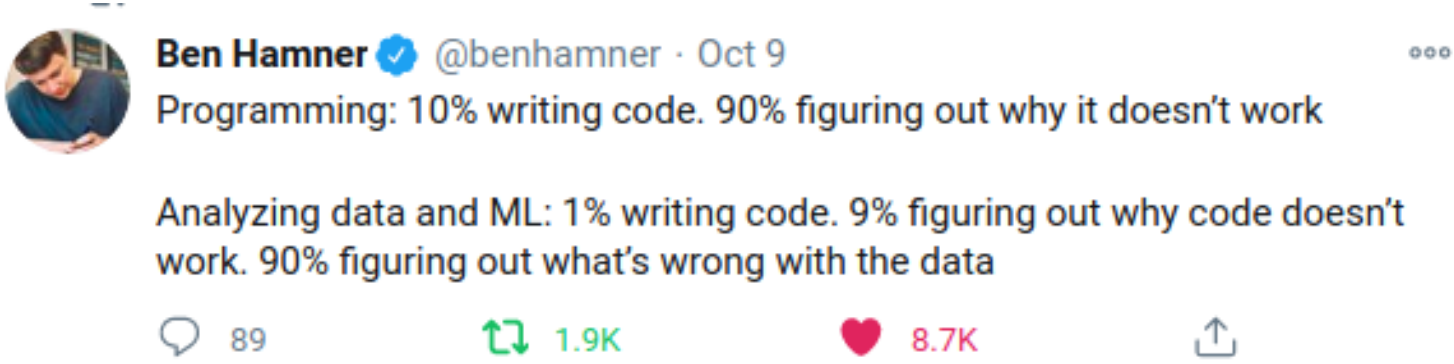
Manual annotation in|for NLP



Manual annotation in|for NLP



Why it's important



Penn Treebank (PTB) [Marcus et al., 1993]

PTB 1:

- POS-taggging correction: ? 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-taggging 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-taggging 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-taggging 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-taggging 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-taggging 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)!

Prague Dependency Treebank [Böhmová et al., 2001]

- 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.

$\Rightarrow$ 5 part-time annotators, 1 senior coordinator, 1 junior coordinator during 1.5 year [Kim et al., 2008]

ESTER

- 100 h of transcribed speech (ESTER evaluation campaign, transcription systems, 2008)
- 1 h of speech = ?

ESTER

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

ESTER

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

$\Rightarrow$ 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

$\rightarrow$ fast evolution of the tools

$\Rightarrow$ annotation should **not** depend on them

Practice

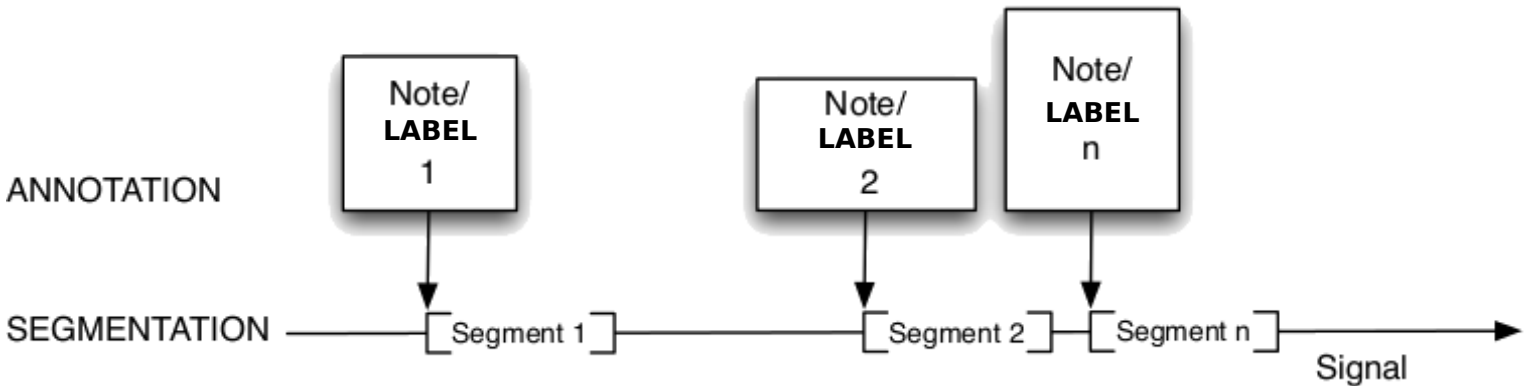
Transcribe what you hear, using any text editor or paper.

Definition

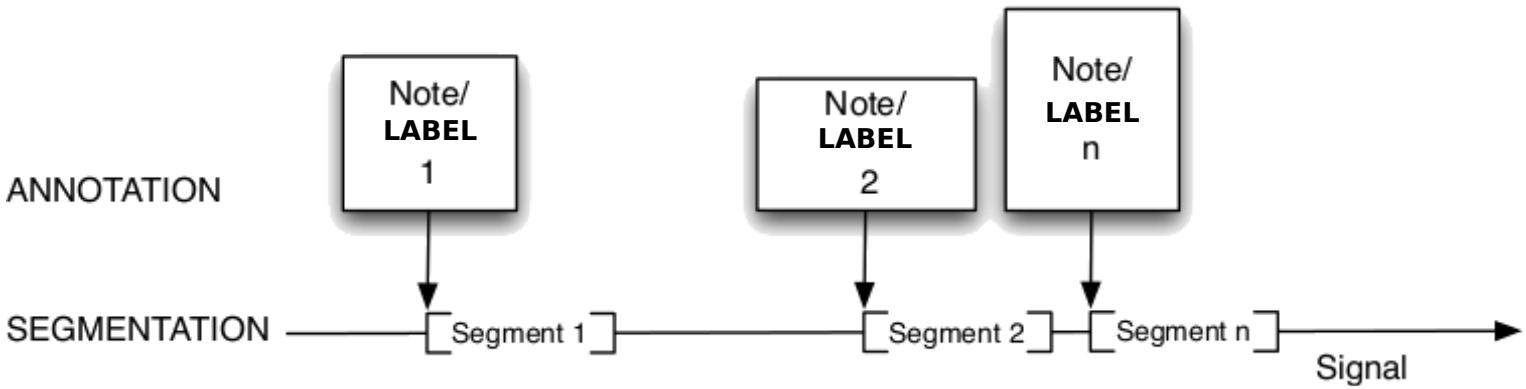
"[corpus annotation] can be defined as the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data. 'Annotation' can also refer to the end-product of this process" [Leech, 1997]

"'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]

Annotation


ANNOTATION
Note/
LABEL
1
Note/
LABEL
2
Note/
LABEL
n
SEGMENTATION
Segment 1
Segment 2
Segment n
Signal

Annotation



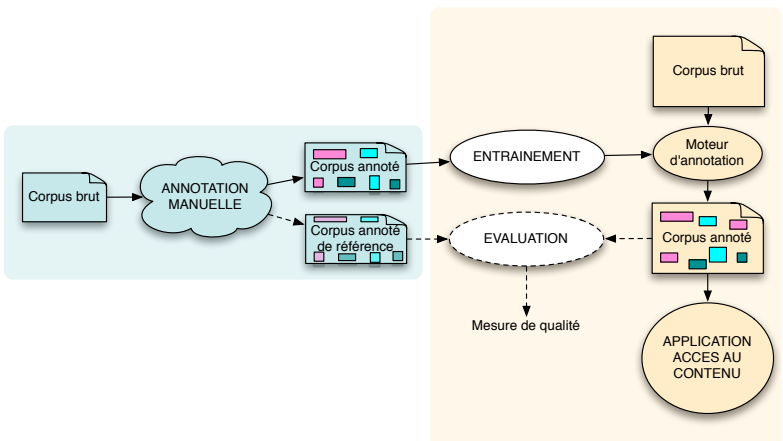
Adding interpretative information [Leech, 1997, Habert, 2005]

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)

[T]he annotations are more useful, the more they are designed to be specific to a particular application [Leech, 2005].



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.

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)

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.)

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

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)

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...)

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)

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

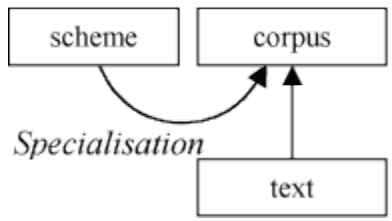
Leech's 7 maxims [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Note: can be hard after normalization ("l'arbre" $\rightarrow$ "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
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]

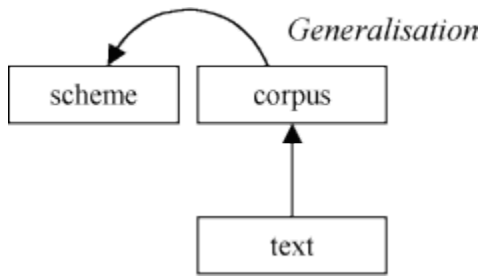
Model-based approach



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

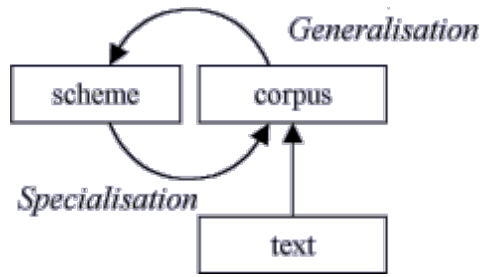
Everything is in the annotation!

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

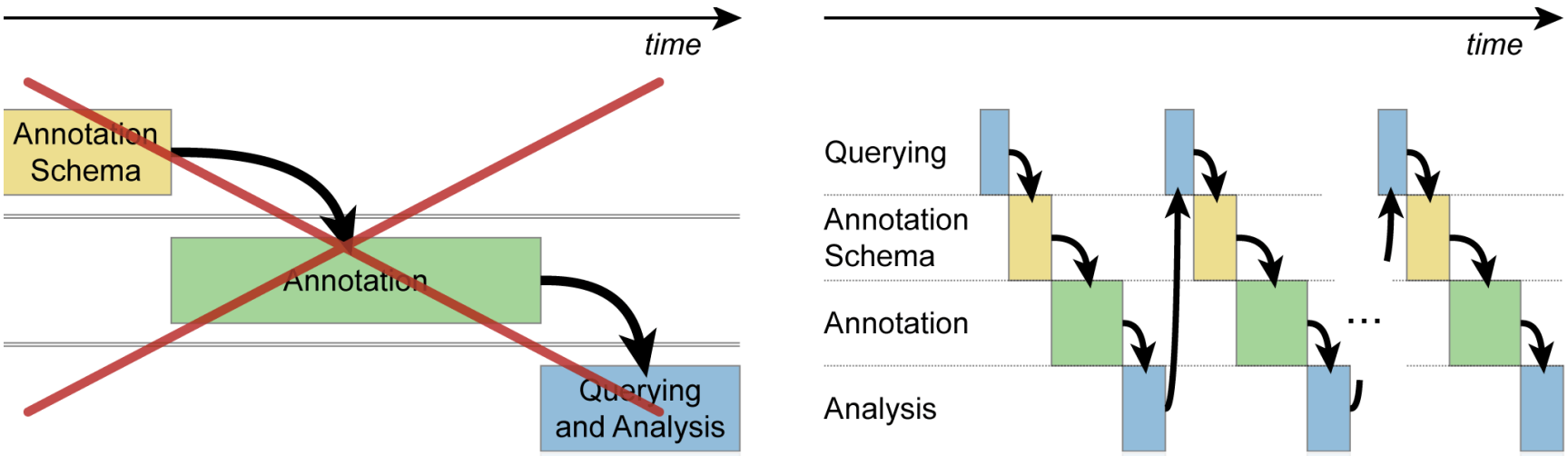
- new observations generalize hypotheses
- theory allows to interpret and classify information

- evolving cycles: each cycle improves the knowledge by refining and testing the theories on real data

$\Rightarrow$ a more precise representation of the corpus is built and a more sophisticated system is produced

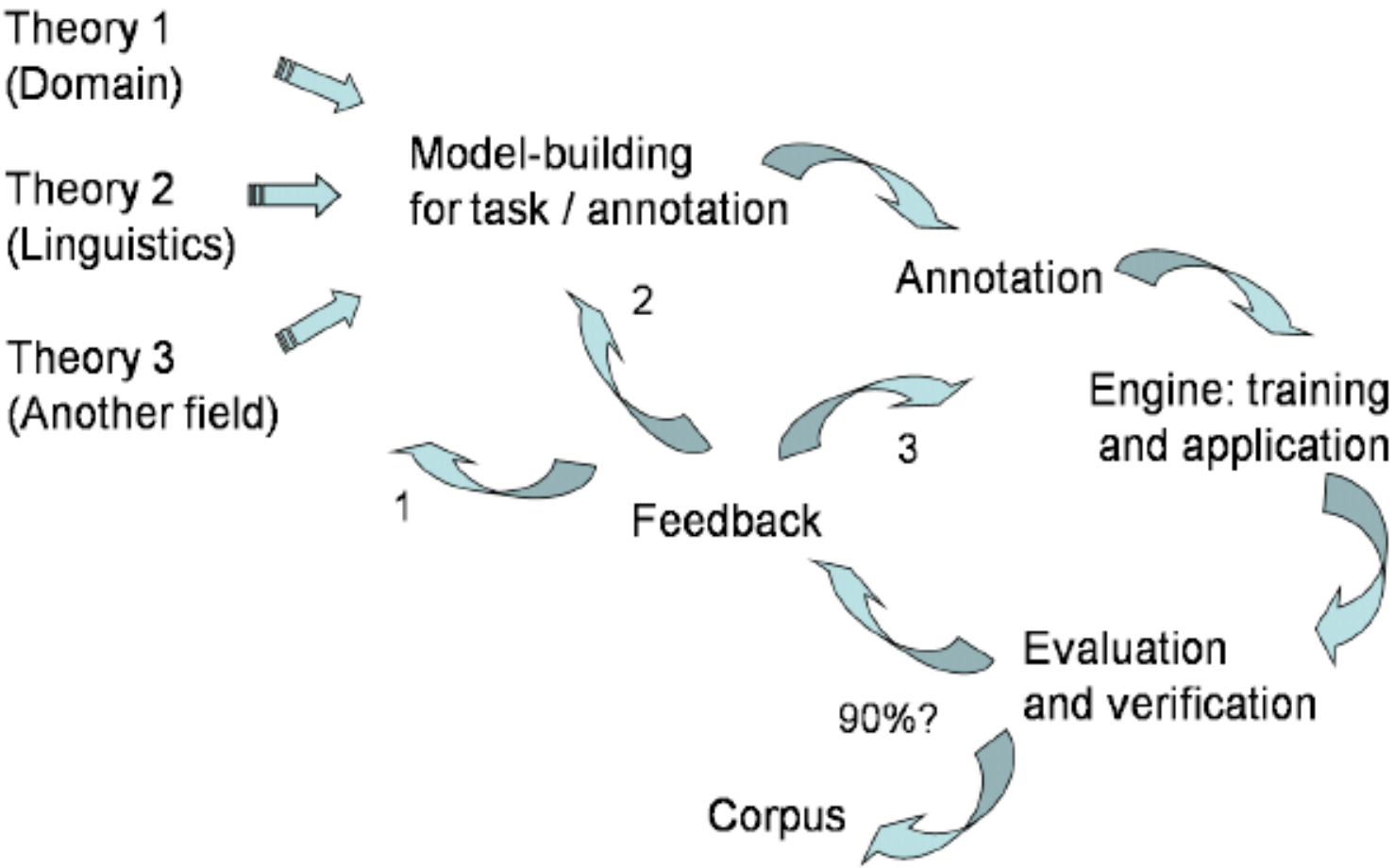
Agile Annotation

integrating evaluation

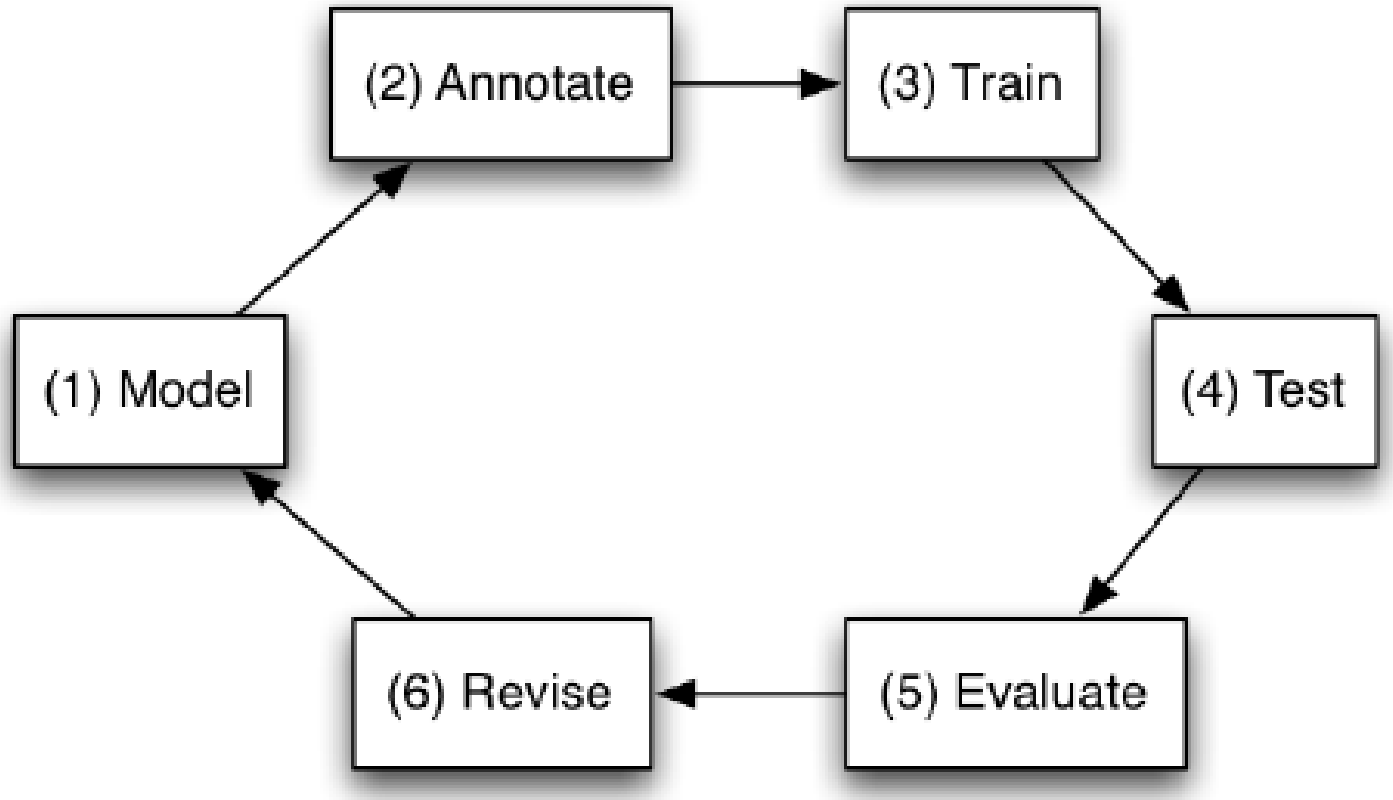


Traditionnal annotation phases (left) and cycles of agile annotation (right). Reproduction of Figure 2 from [Voormann and Gut, 2008]

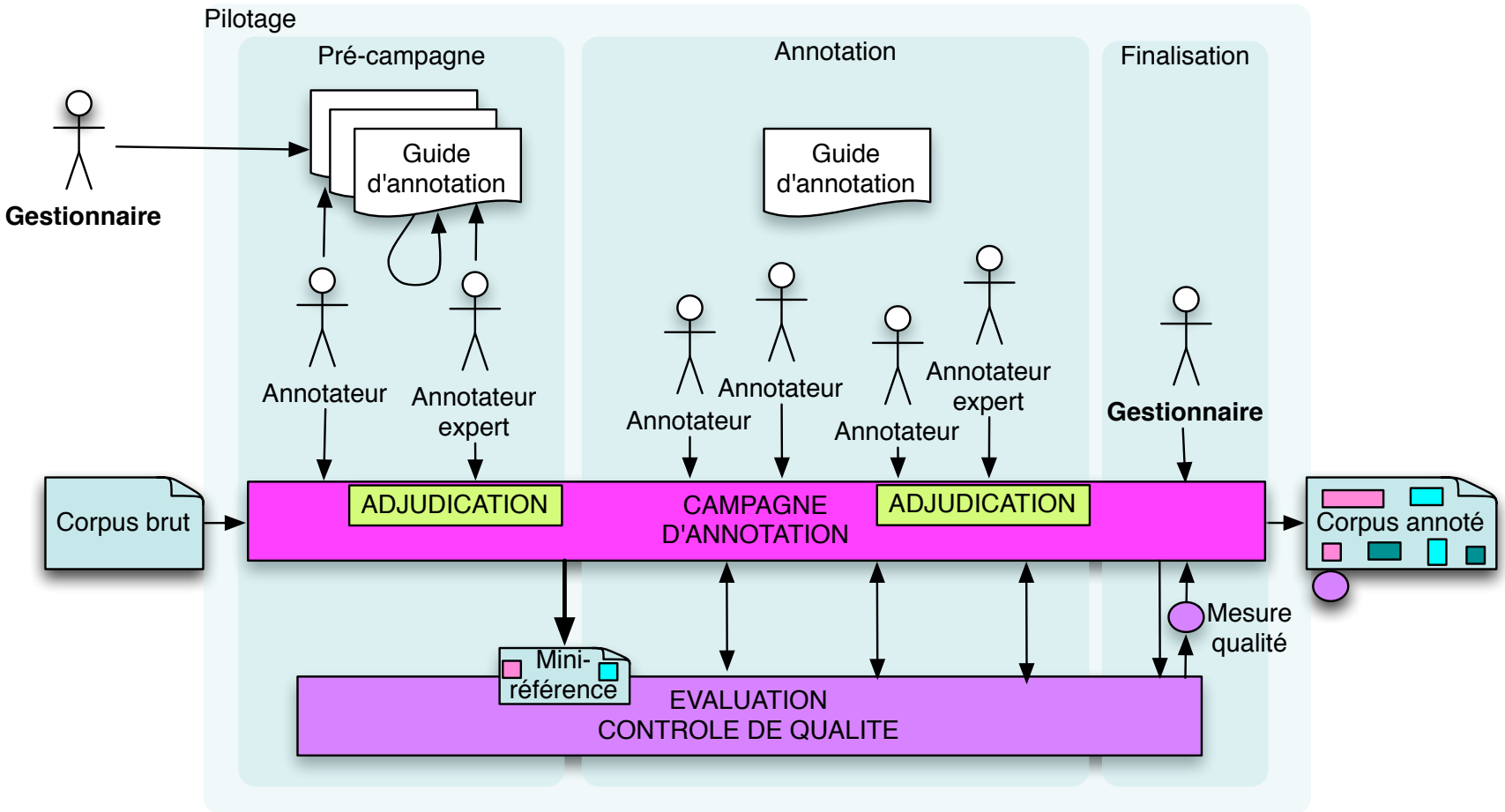
Generic annotation pipeline [Hovy and Lavid, 2010]



MATTER cycle [Pustejovsky and Stubbs, 2012]

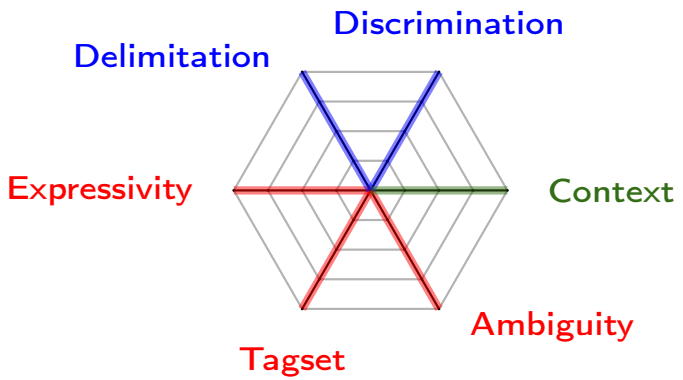


Towards "annotation engineering" [Fort, 2012]



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.

$\Rightarrow$ **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 *should* be annotated as compared to what *could* be annotated is low.

Definition

$$Discrimination(Flow) = 1 - \frac{|Annotations(Flow)|}{\sum_{i=1}^{LevelSeg} |UnitsObtainedBySeg_i(Flow)|}$$

$\Rightarrow$ 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 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.

$$Discrimination_{Identification} = 0,9$$
$$Discrimination_{Renaming} = 0,95$$

Boundaries delimitation

- **extending** or **shrinking** the discriminated unit:
  Madame Chirac $\rightarrow$ *Monsieur et Madame Chirac*

Boundaries delimitation

- **extending** or **shrinking** the discriminated unit:
  Madame Chirac $\rightarrow$ *Monsieur et Madame Chirac*
- **decompose** a discriminated unit into several elements:
  le préfet Érignac $\rightarrow$ *le préfet Érignac*

Boundaries delimitation

- **extending** or **shrinking** the discriminated unit:
  Madame Chirac $\rightarrow$ *Monsieur et Madame Chirac*
- **decompose** a discriminated unit into several elements:
  le préfet Érignac $\rightarrow$ *le préfet Érignac*
- or **group** together several discriminated units into one unique annotation:
  Sa Majesté
  le roi Mohamed VI $\rightarrow$ *Sa Majesté le roi Mohamed VI*

Boundaries delimitation

Definition

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

$$Delimitation_{Identification} = 0$$
$$Delimitation_{Renaming} = 0$$

$$Delimitation_{PTB_{POS}} = 0$$

$$Délimitation_{EN_{TypesSubtypes}} = 1$$
$$Délimitation_{EN_{Components}} = 0,3$$

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

$$Expressiveness_{Identification} = 0.25$$
$$Expressiveness_{Renaming} = 0.25$$

Dimension of the tagset

<table>
<tr><th colspan="2">Person</th><th colspan="2">Function</th></tr>
<tr><td>pers.ind (individual person)</td><td>pers.coll (group of persons)</td><td>func.ind (individual function)</td><td>func.coll (collectivity of functions)</td></tr>
<tr><th colspan="3">Location</th><th colspan="3">Production</th></tr>
<tr><td colspan="3" rowspan="3">administrative (loc.adm.town, loc.adm.reg, loc.adm.nat, loc.adm.sup)
physical (loc.phys.geo, loc.phys.hydro, loc.phys.astro)
facilities (loc.fac), oronyms (loc.oro), address (loc.add.phys, loc.add.elec)</td><td>prod.object (manufac-tured object)</td><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><td>prod.soft (software)</td></tr>
<tr><td>prod.art</td><td>prod.media</td><td>prod.award</td></tr>
<tr><th colspan="2">Organization</th><th colspan="2" rowspan="2">Time</th></tr>
<tr><td>org.adm (administra-tion)</td><td>org.ent (services)</td></tr>
<tr><th colspan="2">Amount</th><td colspan="2" rowspan="2">time.date.abs (absolute date), time.date.rel (relative date)</td><td colspan="2" rowspan="2">time.hour.abs (absolute hour), time.hour.rel (relative hour)</td></tr>
<tr><td colspan="2">amount (with unit or general object), includ-ing duration</td></tr>
</table>

Types and sub-types used for structured NE annotation [Grouin et al., 2011]

Dimension of the tagset

<table>
<tr><td colspan="2">Person</td><td colspan="2">Function</td></tr>
<tr><td>pers.ind (individual person)</td><td>pers.coll (group of persons)</td><td>func.ind (individual function)</td><td>func.coll (collectivity of functions)</td></tr>
<tr><td colspan="2">Location</td><td colspan="2">Production</td></tr>
<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><td colspan="2">facilities (loc.fac), oronyms (loc.oro), address (loc.add.phys, loc.add.elec)</td></tr>
<tr><td colspan="2">Organization</td><td colspan="2">Time</td></tr>
<tr><td>org.adm (administration)</td><td>org.ent (services)</td><td colspan="2" rowspan="2">time.date.abs (absolute date), time.date.rel (relative date)</td></tr>
<tr><td colspan="2">Amount</td></tr>
<tr><td colspan="2">amount (with unit or general object), including duration</td><td colspan="2">time.hour.abs (absolute hour), time.hour.rel (relative hour)</td></tr>
</table>

<table>
<tr><td>prod.object (manufactured object)</td><td>prod.serv (transportation route)</td><td>prod.fin (financial products)</td></tr>
<tr><td>prod.doctr (doctrine)</td><td>prod.rule (law)</td><td>prod.soft (software)</td></tr>
<tr><td>prod.art</td><td>prod.media</td><td>prod.award</td></tr>
</table>

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

Dimension of the tagset

Person		**Function**	
pers.ind (individual person)	*pers.coll* (group of persons)	*func.ind* (individual function)	*func.coll* (collectivity of functions)

Location			**Production**		
administrative (*loc.adm.town*, *loc.adm.reg*, *loc.adm.nat*, *loc.adm.sup*)	physical (*loc.phys.geo*, *loc.phys.hydro*, *loc.phys.astro*)	facilities (*loc.fac*), oronyms (*loc.oro*), address (*loc.add.phys*, *loc.add.elec*)	*prod.object* (manufactured object)	*prod.serv* (transportation route)	*prod.fin* (financial products)
			prod.doctr (doctrine)	*prod.rule* (law)	*prod.soft* (software)
			prod.art	*prod.media*	*prod.award*

Organization		**Time**	
org.adm (administration)	*org.ent* (services)	*time.date.abs* (absolute date), *time.date.rel* (relative date)	*time.hour.abs* (absolute hour), *time.hour.rel* (relative hour)
Amount			
amount (with unit or general object), including duration			

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).

Dimension of the tagset

<table>
<tr><th colspan="3">Person</th><th colspan="3">Function</th></tr>
<tr><td colspan="2">*pers.ind* (individual person)</td><td>*pers.coll* (group of persons)</td><td colspan="2">*func.ind* (individual function)</td><td>*func.coll* (collectivity of functions)</td></tr>
<tr><th colspan="3">Location</th><th colspan="3">Production</th></tr>
<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><td>facilities (*loc.fac*), oronyms (*loc.oro*), address (*loc.add.phys, loc.add.elec*)</td><td>*prod.object* (manufac-tured object)</td><td>*prod.serv* (transporta-tion route)</td><td>*prod.fin* (financial products)</td></tr>
<tr><td colspan="3" rowspan="3"></td><td>*prod.doctr* (doctrine)</td><td>*prod.rule* (law)</td><td>*prod.soft* (software)</td></tr>
<tr><td>*prod.art*</td><td>*prod.media*</td><td>*prod.award*</td></tr>
<tr><td colspan="3"></td></tr>
<tr><th colspan="2">Organization</th><td></td><th colspan="2">Time</th><td></td></tr>
<tr><td colspan="2">*org.adm* (administra-tion)</td><td>*org.ent* (services)</td><td colspan="2" rowspan="2">*time.date.abs* (absolute date), *time.date.rel* (relative date)</td><td rowspan="2">*time.hour.abs* (absolute hour), *time.hour.rel* (relative hour)</td></tr>
<tr><th colspan="3">Amount</th></tr>
<tr><td colspan="3">*amount* (with unit or general object), includ-ing duration</td><td colspan="3"></td></tr>
</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: *loc.adm.town*, *loc.adm.reg*, *loc.adm.nat*, *loc.adm.sup* $\rightarrow$ 4 possibilities (degree of freedom = 3).

Dimension of the tagset

Degree of freedom

$$\nu = \nu_1 + \nu_2 + \ldots + \nu_m$$

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

$$Dimension_{Identification} = 0$$
$$Dimension_{Renaming} = 0.04$$
$$Dimension_{NE_{TypesSubtypes}} = 0.34$$

Degree of ambiguity: residual ambiguity

Using the traces left by the annotators:

[...] <EukVirus>3CDproM</EukVirus> can process both structural and nonstructural precursors of the <EukVirus ***uncertainty-type*** *= "too-generic"><taxon>poliovirus</taxon> polyprotein</EukVirus> [...].*

Définition

$$AmbiguityRes(Flow) = \frac{|Annotations_{amb}|}{|Annotations|}$$

$$AmbiguityRes_{Identification} = 0.04$$
$$AmbiguityRes_{Renaming} = 0.02$$

Degree of ambiguity: theoretical ambiguity

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

Definition

$$AmbiguityTh(Flow) = \frac{\sum_{voc_i=1}^{|Voc(Flow)|} (Ambig(voc_i) * freq(voc_i, Flow))}{|Units(Flow)|}$$

with

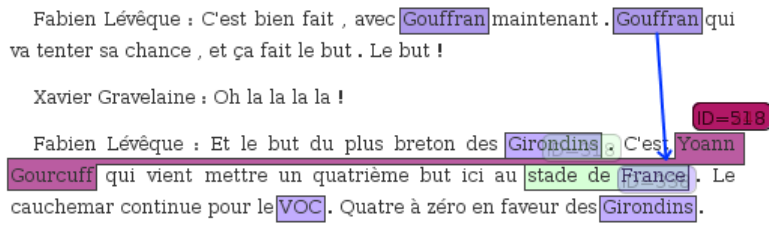
$$Ambig(voc_i) = \begin{cases} 1 & \text{if} \quad |Tags(voc_i)| > 1 \\ 0 & \text{else} \end{cases}$$

$$AmbiguityTh_{Identification} = 0.01$$

$\rightarrow$ Does not apply to renaming relations

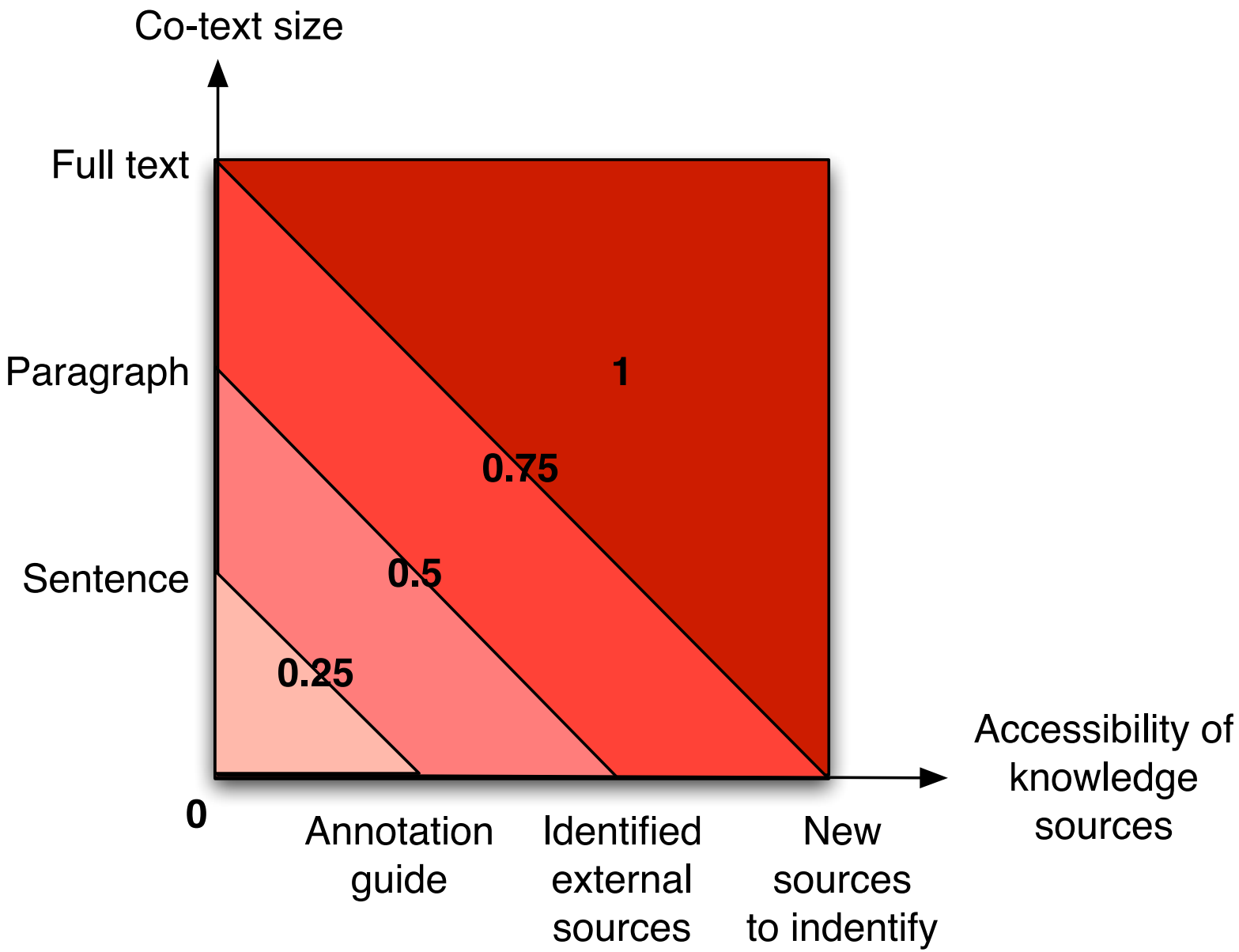
Context to take into account

- **size of the window** to take into account in the source signal:
  - The sentence:
    I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.
  - ... or more:

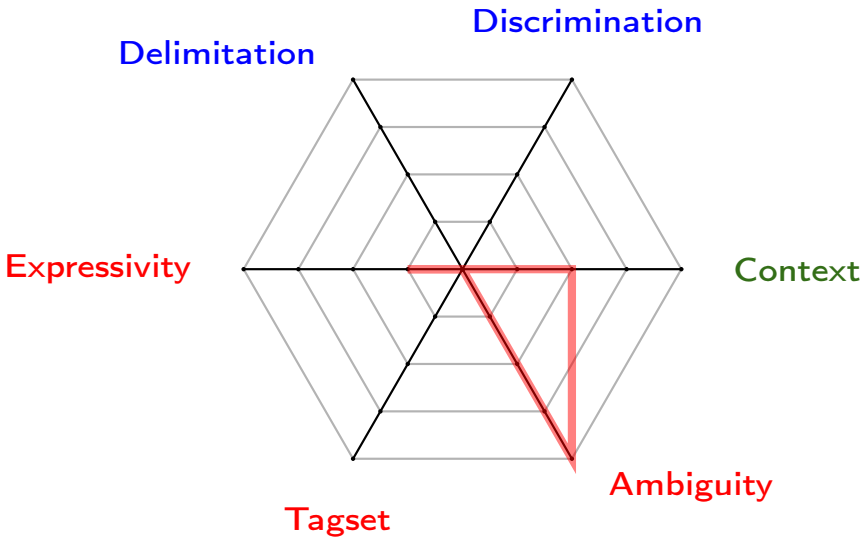


- number of **knowledge elements** to be rallied or degree of accessibility of the knowledge sources that are consulted:
  - annotation guidelines
  - nomenclatures (Swiss-Prot)
  - new sources to be found (Wikipedia, etc.)

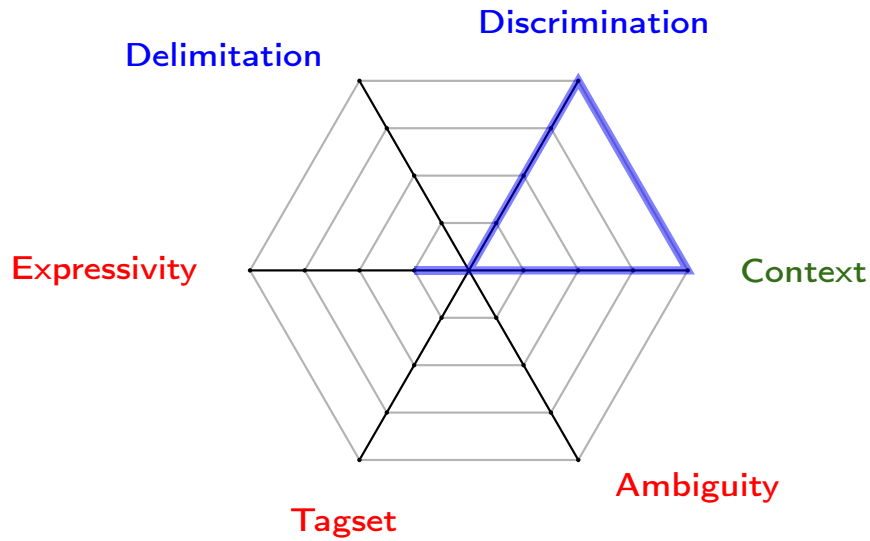
Weight of the context



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

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
International Conference on Computational Linguistics, Dec 2012, Mumbaï, India. pp.895–910
https://hal.archives-ouvertes.fr/hal-00769631/file/coling2012_Complexity_KF_30102012.pdf


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


Böhmová, A., Hajič, J., Hajičová, E., and Hladká, B. (2001).
The prague dependency treebank: Three-level annotation scenario.
In Abeillé, A., editor, Treebanks: Building and Using Syntactically Annotated Corpora. Kluwer Academic Publishers.


Church, K. W. (1988).
A stochastic parts program and noun phrase parser for unrestricted text.
In Proceedings of the Second Conference on Applied Natural Language Processing, ANLC '88, pages 136–143, Stroudsburg, PA, USA. Association for Computational Linguistics.


Desrosières, A. (2008).
Pour une sociologie historique de la quantification : L'Argument statistique.

Presses de l'école des Mines de Paris.


Fort, K. (2012).
Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus.
PhD thesis, Université Paris XIII, LIPN, INIST-CNRS.


Fort, K., François, C., Galibert, O., and Ghribi, M. (2012).
Analyzing the impact of prevalence on the evaluation of a manual annotation campaign.
In Proceedings of the International Conference on Language Resources and Evaluation (LREC), Istanbul, Turquie.
7 pages.


Grouin, C., Rosset, S., Zweigenbaum, P., Fort, K., Galibert, O., and Quintard, L. (2011).
Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview.
In Proceedings of the 5th Linguistic Annotation Workshop, pages 92–100, Portland, Oregon, USA.
Poster.

Habert, B. (2000).
Corpus. Méthodologie et applications linguistiques, chapter Détournements d'annotation : armer la main et le regard, pages 106–120.
Champion and Presses Universitaires de Perpignan.

Habert, B. (2005).
Portrait de linguiste(s) à l'instrument.
Texto!, vol. X(4).

Hovy, E. H. and Lavid, J. M. (2010).
Towards a "science" of corpus annotation: A new methodological challenge for corpus linguistics.
International Journal of Translation Studies, 22(1).

Kim, J.-D., Ohta, T., and Tsujii, J. (2008).
Corpus annotation for mining biomedical events from literature.

BMC Bioinformatics, 9(1):10.

Leech, G. (1993).

Corpus annotation schemes.
Literary and Linguistic Computing, 8(4):275–281.


Leech, G. (1997).
Corpus annotation: Linguistic information from computer text corpora, chapter Introducing corpus annotation, pages 1–18.
Longman, Londres, Angleterre.


Leech, G. (2005).
Developing Linguistic Corpora: a Guide to Good Practice, chapter Adding Linguistic Annotation, pages 17–29.
Oxford: Oxbow Books.


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.


Pustejovsky, J. and Stubbs, A. (2012).
Natural Language Annotation for Machine Learning.
O'Reilly.


Voormann, H. and Gut, U. (2008).
Agile corpus creation.
Corpus Linguistics and Linguistic Theory, 4(2):235–251.


Wallis, S. (2007).
Annotating Variation and Change, chapter Annotation, Retrieval and Experimentation.
Varieng, University of Helsinki, Helsinki, Finland.