CC BY NC SA

Manual Annotation in NLP: Motivations and definitions

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

Today's NLP
NLP: applications in our lives
NLP: applications in our work
Language resources remain at the heart of NLP

Manual annotation and NLP

What is annotating about ?

How to do this properly?

WYMR: What You Must Remember

Today's NLP
NLP: applications in our lives
NLP: applications in our work
Language resources remain at the heart of NLP

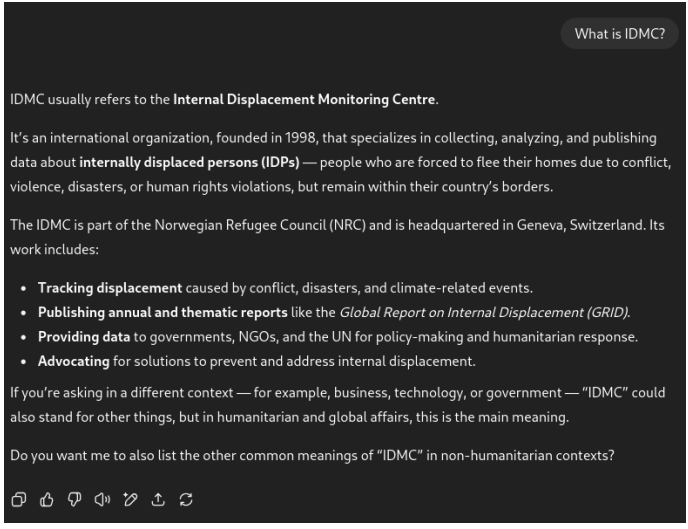
Manual annotation and NLP

What is annotating about ?

How to do this properly?

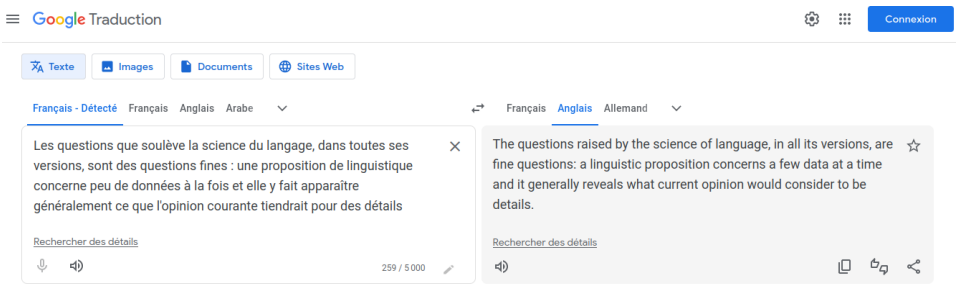
WYMR: What You Must Remember

ChatBot (usage?)



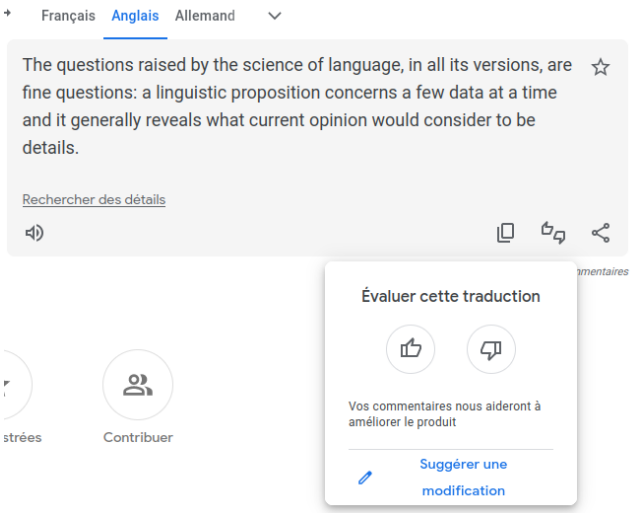
https://chatgpt.com/

Machine Translation



https://translate.google.com/

Machine Translation



https://translate.google.com/

Today's NLP
NLP: applications in our lives
NLP: applications in our work
Language resources remain at the heart of NLP

Manual annotation and NLP

What is annotating about ?

How to do this properly?

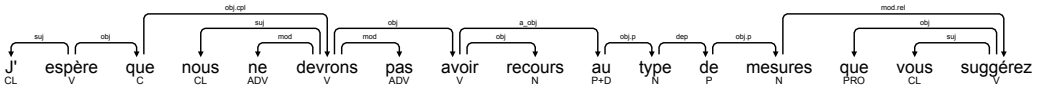
WYMR: What You Must Remember

Example: Part-of-Speech tagging

('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 .

Example: Dependency syntax



Example: Named Entity Extraction

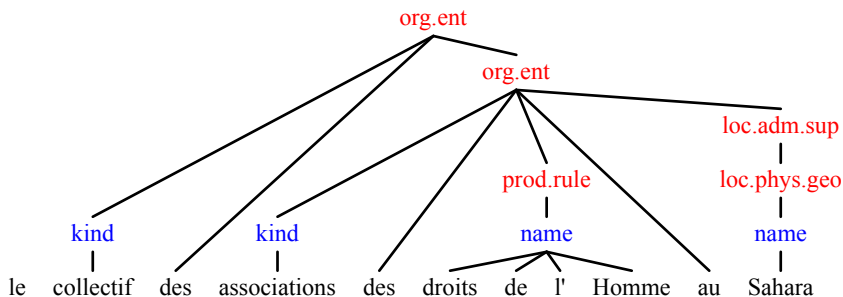
<ENAMEX TYPE="PERSON">*Henri****</ENAMEX>*** *a acheté* ***<NUMEX TYPE="QUANTITY">****300****</NUMEX>*** *actions de la société* ***<ENAMEX TYPE="ORGANIZATION">****AMD****</ENAMEX>*** *en* ***<TIMEX TYPE="DATE">****2006****</TIMEX>****.*

http://fr.wikipedia.org/wiki/Reconnaissance_d%27entit%C3%A9s_nomm%C3%A9es

Example: Named Entity Extraction

<ENAMEX TYPE="PERSON">Henri</ENAMEX> a acheté <NUMEX TYPE="QUANTITY">300</NUMEX> actions de la société <ENAMEX TYPE="ORGANIZATION">AMD</ENAMEX> en <TIMEX TYPE="DATE">2006</TIMEX>.

http://fr.wikipedia.org/wiki/Reconnaissance_d%27entit%C3%A9s_nomm%C3%A9es



[Grouin et al., 2011]

Today's NLP
NLP: applications in our lives
NLP: applications in our work
Language resources remain at the heart of NLP

Manual annotation and NLP

What is annotating about ?

How to do this properly?

WYMR: What You Must Remember

Language resources: at the heart of NLP

- systems based on data (99.9% today)
  - (un)supervised machine learning (deep or not)
  - from examples (written and/or annotated by humans)
  - neural or statistical algorithms (created by humans)

  → raw and annotated corpora

[Dependency Syntax Annotation]

- systems based on rules (used to be the 99%)
  - defined by humans (usually linguists)
  - manuelly entered/coded

  → grammars, lexicons

```
mirador         nc      [pred='mirador_____1<Suj:(sn),Objde:(de-sn|de-sinf),Objà:(à-sinf)>',cat=nc,@ms] mirador_____1   Default ms
miradors        nc      [pred='mirador_____1<Suj:(sn),Objde:(de-sn|de-sinf),Objà:(à-sinf)>',cat=nc,@mp] mirador_____1   Default mp
mirage          nc      [pred='mirage_____1<Suj:(sn),Objde:(de-sn|de-sinf),Objà:(à-sinf)>',cat=nc,@ms]  mirage_____1    Default ms
mirages         nc      [pred='mirage_____1<Suj:(sn),Objde:(de-sn|de-sinf),Objà:(à-sinf)>',cat=nc,@mp]  mirage_____1    Default mp
```

[Le*fff*, [Sagot, 2010]]

Today's NLP

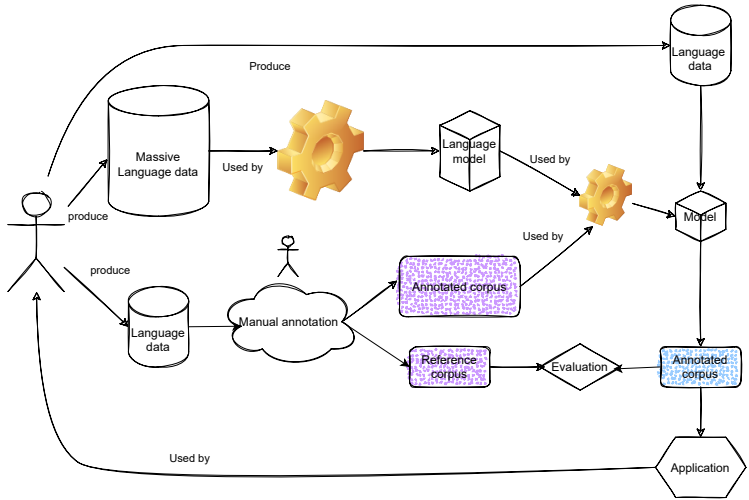
Manual annotation and NLP
- Manual annotation in NLP
- A notoriously costly endeavour
- About language resources longevity

What is annotating about ?

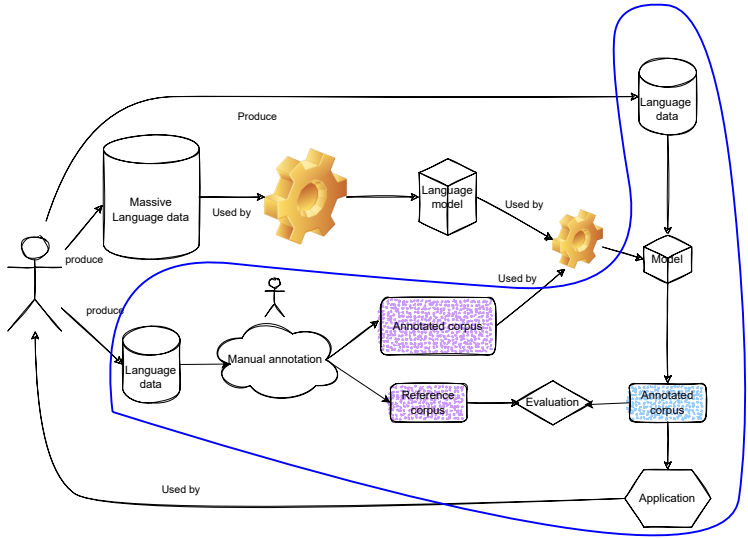
How to do this properly?

WYMR: What You Must Remember

Today's NLP


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

Today's NLP


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

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

89 1.9K 8.7K

Why it's important? in real life

"Manual inspection of errors in the CIC-ES and CIC-CAT datasets revealed that a substantial portion of misclassifications were due to dataset quality issues rather than model shortcomings. In our annotated sample, approximately 60% of errors in both languages were attributable to incorrect labels or tweets that were unrelated to the topic of Catalan independence. [...]"

"These findings highlight a critical point: much of the perceived model underperformance may be due to label inconsistencies and annotation errors. In fact, the evaluated LLMs may be classifying better than the raw scores indicate. However, confirming this hypothesis would require systematic re-annotation or cleaning of the CIC dataset. " [Masters' thesis of Summer Devlin, M2 student in 2025]

Today's NLP

Manual annotation and NLP
- Manual annotation in NLP
- A notoriously costly endeavour
- About language resources longevity

What is annotating about ?

How to do this properly?

WYMR: What You Must Remember

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

Treebank from the University of Pennsylvanie

PTB 1:

- correcting automatic POS-tagging: ? words per hour, ? hours a day
- correcting automatic parsing: ? words per hour, ? hours a day

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

Treebank from the University of Pennsylvanie

PTB 1:

- correcting automatic POS-tagging: 3,000 words per hour, ? hours a day
- correcting automatic parsing: ? words per hour, ? hours a day

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

Treebank from the University of Pennsylvanie

PTB 1:

- correcting automatic POS-tagging: 3,000 words per hour, 3 hours a day
- correcting automatic parsing: ? words per hour, ? hours a day

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

Treebank from the University of Pennsylvanie

PTB 1:

- correcting automatic POS-tagging: 3,000 words per hour, 3 hours a day
- correcting automatic parsing: 750 words per hour, ? hours a day

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

Treebank from the University of Pennsylvanie

PTB 1:

- correcting automatic POS-tagging: 3,000 words per hour, 3 hours a day
- correcting automatic parsing: 750 words per hour, 3 hours a day

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

Treebank from the University of Pennsylvanie

PTB 1:

- correcting automatic POS-tagging: 3,000 words per hour, 3 hours a day
- correcting automatic parsing: 750 words per hour, 3 hours a day

\+ learning curve of 1 (POS-tagging) to 2 months (syntax)!

Prague Dependency Treebank [Böhmová et al., 2001]

corpus annotated in dependency syntax from Charles' University

- 1996-2004 [Böhmová et al., 2001],
- built from the CNC (*Czech National Corpus*),
- 3 levels of structure:
  1. morphological (semi-automatic): 1.8 million tokens
  2. analytical (dependency syntax, with an adapted tool)
  3. tectogrammatical (semantic): 1 million tokens

Prague Dependency Treebank [Böhmová et al., 2001]

Version 1.0:

- manual annotation of the morpholocal and analytical levels
- time: ?
- nb of persons involved: ?
- Cost estimate: ?

Prague Dependency Treebank [Böhmová et al., 2001]

Version 1.0:

- manual annotation of the morpholocal and analytical levels
- time: 5 years
- nb of persons involved: ?
- Cost estimate: ?

Prague Dependency Treebank [Böhmová et al., 2001]

Version 1.0:

- manual annotation of the morpholocal and analytical levels
- time: 5 years
- nb of persons involved: 22 persons, incl. 17 simultaneously during pick periods
- Cost estimate: ?

Prague Dependency Treebank [Böhmová et al., 2001]

Version 1.0:

- manual annotation of the morpholocal and analytical levels
- time: 5 years
- nb of persons involved: 22 persons, incl. 17 simultaneously during pick 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 half-time annotators, 1 senior coordinator, 1 junior coordinator during 1.5 year [Kim et al., 2008]

GENIA [Kim et al., 2008]

GENIA: 400,000 words annotated in microbiology.

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

$\Rightarrow$ Quality must be high!

ESTER

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

ESTER

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

Today's NLP

Manual annotation and NLP
- Manual annotation in NLP
- A notoriously costly endeavour
- About language resources longevity

What is annotating about ?

How to do this properly?

WYMR: What You Must Remember

Lifespan of annotated corpora

Penn Treebank [Marcus et al., 1993] :

- created at the beginning of the 90s
- still used (ACL 2022)

vs PARTS POS-tagger [Church, 1988], used to pre-annotate the corpus, which is not more used or even known

$\rightarrow$ rapid evolution of tools

$\Rightarrow$ manual annotation should **not** depend on them/their performance

LLM as a judge?

- What is it?
- Can it replace human annotation?

Today's NLP

Manual annotation and NLP

What is annotating about ?
- Exercise
- DefinitionS
- Annotating, what for?

How to do this properly?

WYMR: What You Must Remember

Exercise

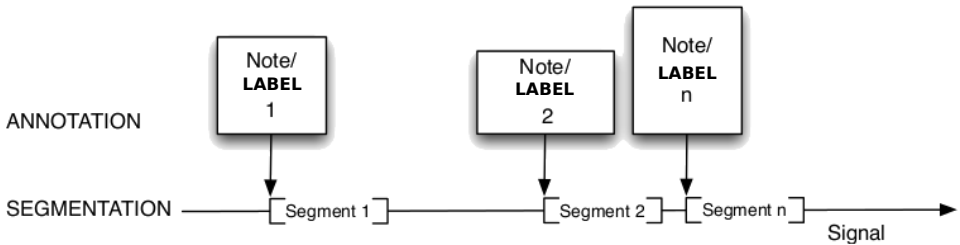
Transcribe what you hear (the file I'll be playing), using Praat (if you have it installed) or on a simple text file, or even on paper (yes)

Definition

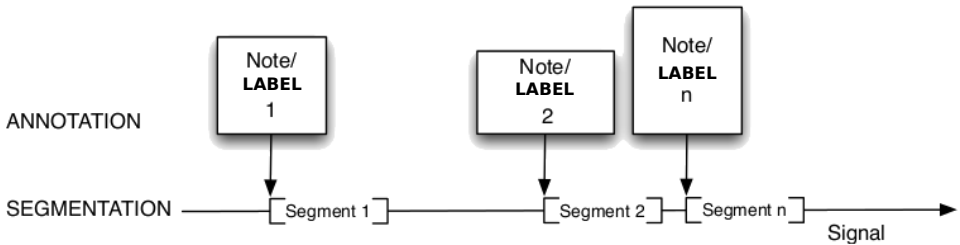
"[corpus annotation] can be defined as the practice of adding interpretative, linguistic information to an electronic corpus of spoken and/or written language data. 'Annotation' can also refer to the end-product of this process" [Leech, 1997]

"'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


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

Annotation



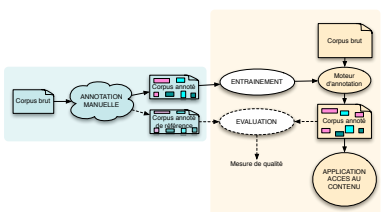
Adding interpretative information [Leech, 1997, Habert, 2005]

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)

Today's NLP

Manual annotation and NLP

What is annotating about ?

How to do this properly?
- Good Practises
- Theorizing

WYMR: What You Must Remember

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

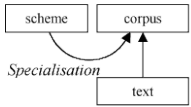
Leech's 7 maxims [Leech, 1993]

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]

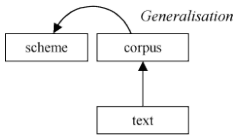
Model-based approach



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

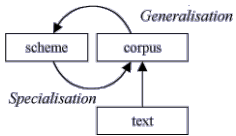
Everything is in the annotation!

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

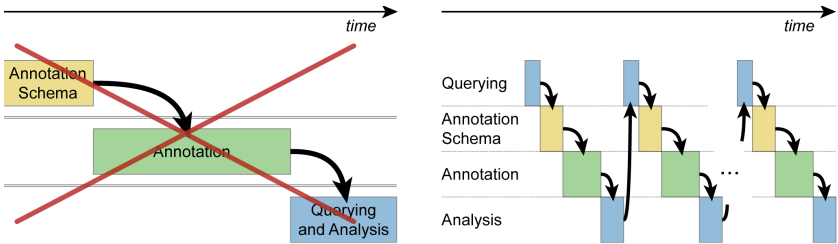
- new observations generalize hypotheses
- theory allows to interpret and classify information

- 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

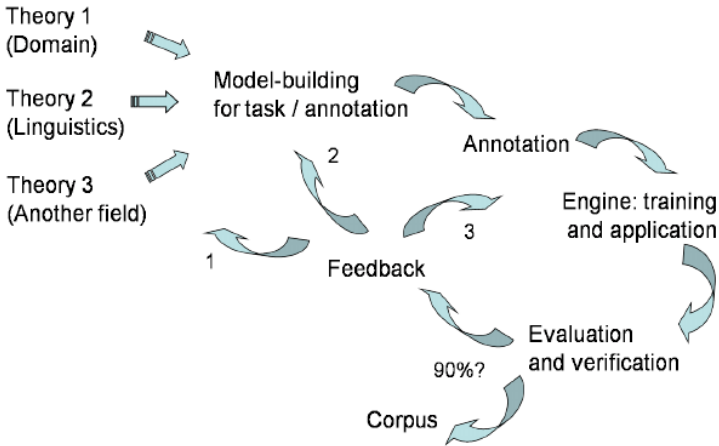
Agile Annotation

integrating evaluation

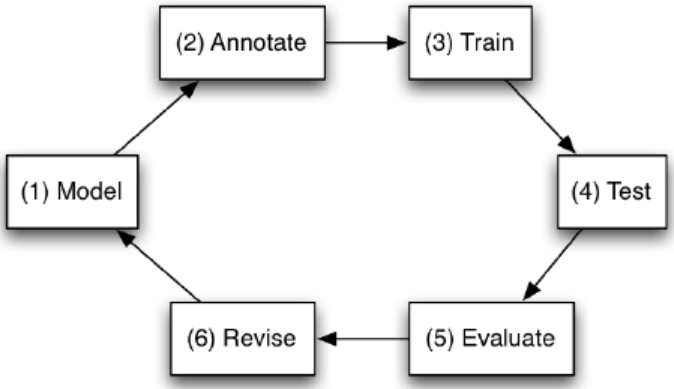


Traditionnal annotation phases (left) and cycles of agile annotation (right).
Reproduction of Figure 2 from [Voormann and Gut, 2008]

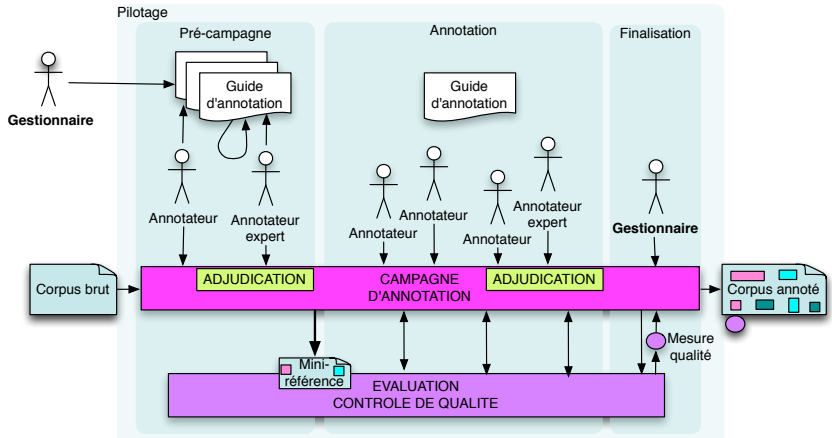
Generic annotation pipeline [Hovy and Lavid, 2010]



MATTER cycle [Pustejovsky and Stubbs, 2012]



Towards "annotation engineering" [Fort, 2012]



Methodology: some basic principles

- compute the inter-annotator agreement at the very beginning of the annotation campaign, then update the annotation guidelines [Bonneau-Maynard et al., 2005].
- compute the intra-annotator agreement as the campaign unfolds, to check that the annotators annotate consistently [Gut and Bayerl, 2004].
- go as far as doing agile annotation [Voormann and Gut, 2008, Alex et al., 2010], that implies several iterations

Manual annotation and NLP:

- usage / importance
- cost

Manual annotation:

- definition
- it's an interpretation

Practice: Manually annotate your corpus with POS

- Which tagset to use?
- Which tool?
- Who annotates what?
- How much time does it take you?
- How do you evaluate the quality of your annotations?

Alex, B., Grover, C., Shen, R., and Kabadjov, M. (2010).
Agile corpus annotation in practice: An overview of manual and automatic annotation of CVs.
In Proceedings of the Fourth Linguistic Annotation Workshop (LAW), pages 29–37, Uppsala, Suède. Association for Computational Linguistics.

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.

Bonneau-Maynard, H., Rosset, S., Ayache, C., Kuhn, A., and Mostefa, D. (2005).
Semantic annotation of the French Media dialog corpus.
In Proceedings of the InterSpeech, Lisbonne, Portugal.

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.

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.

Gut, U. and Bayerl, P. S. (2004).
Measuring the reliability of manual annotations of speech corpora.
In Proceedings of the Speech Prosody, pages 565–568, Nara, Japon.

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.

Sagot, B. (2010).
The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French.
In
7th international conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta.

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.