

Corpus Annotation for Natural Language Processing

Thierry Poibeau (CNRS & PSL/ENS, Lattice)

thierry.poibeau@ens.psl.eu

Oct 23, 2024

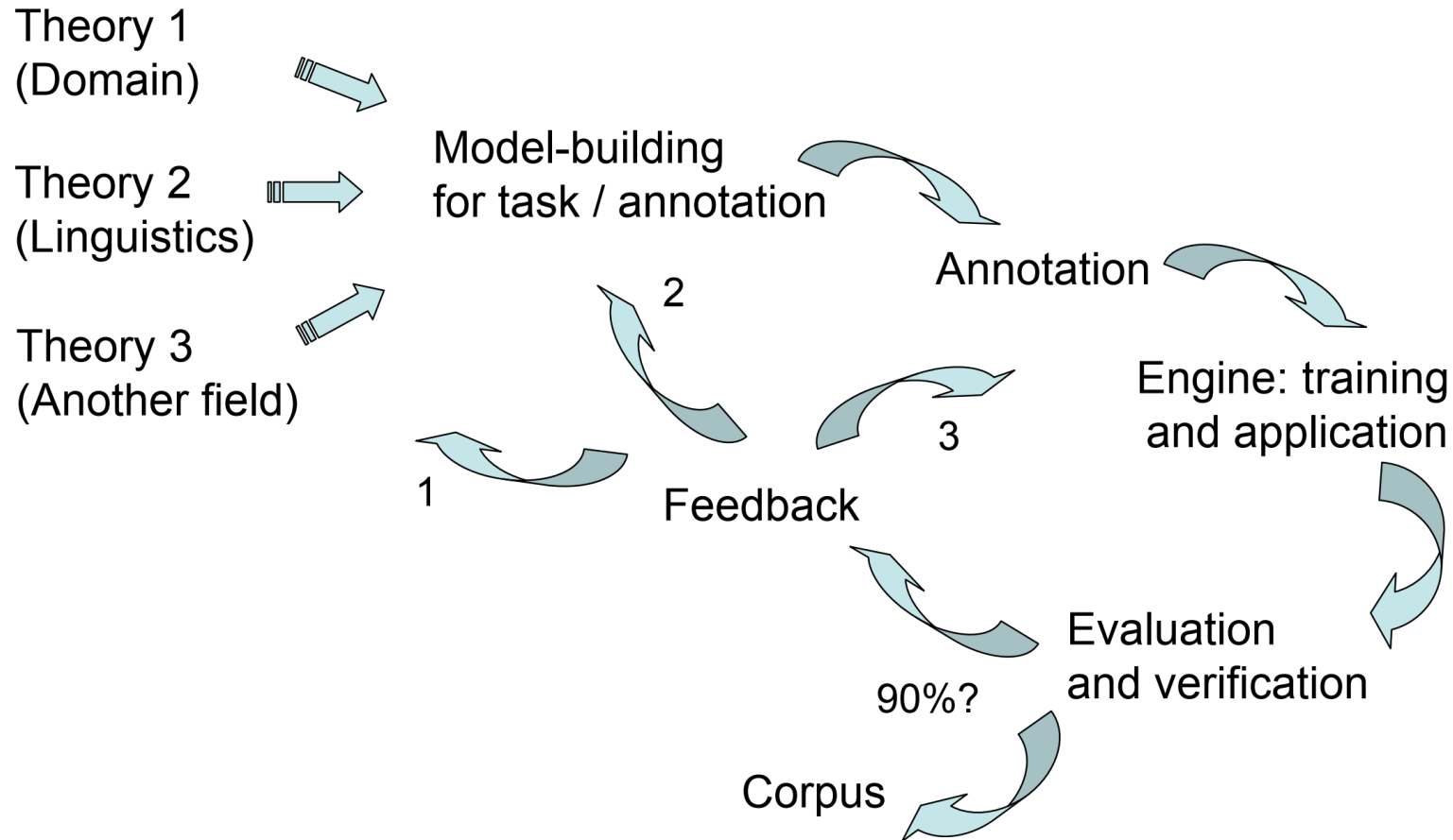
- Based on a series of slides from E. Hovy, a tutorial on text annotation

The Need for Annotated Corpora

The Need for Annotated Corpora

- The ‘machine learning revolution’ in speech and NLP is now complete
 - Most people see speech and NL processing as a notation rewrite problem:
Speech → text, Italian → Chinese, sentence → parse tree, long text → short text...
 - This is the most powerful and effective approach
 - But most approaches are supervised, i.e. they need a (large number) of annotated examples
- Results
 - A new hunger for annotated corpora
 - A new class of researcher: the Annotation Expert
 - BUT: How rigorous is Annotation as a ‘science’?

The generic annotation pipeline



Annotation project desiderata

- Annotation must be
 - Fast... to produce enough material
 - Consistent... enough to support learning
 - Deep... enough to be interesting
- Thus, need
 - Simple procedure and good interface
 - Several people for cross-checking
 - Careful attention to the source theory / the research goal!

Annotation as a science

- Increased need for corpora and for annotation raise new questions
 - What kinds/aspects of ‘domain semantics’ to annotate?
 - ...it’s hardly an uncontroversial notion...
 - Which corpora? How much?
 - Which computational tools to apply once annotation is ‘complete’? When is it complete?
 - How to manage the whole process?
- Need to systematize annotation process — BUT: How rigorous is Annotation as a ‘science’?

Semantic annotation project: BookNLP

- BookNLP project (D. Bamman, Berkeley, 2014-2020)
 - Purpose : annotation of literary corpora (for literary studies and cultural analytics studies)
 - English / multilingual (Spanish, Japanese, Russian, German)
 - In constant evolution: CRF → LSTM → Deep learning
- Annotations
 - Entity annotations (sequences)
 - Coreference annotations (relations)
 - Event annotations (sequences)
 - Quotation annotations (sequences + relations)

The Annotation Process

Seven steps to heaven

- T1: Selecting a corpus
- T2: Instantiating the theory
- T3: Designing the interface
- T4: Selecting and training the annotators
- T5: Designing and managing the annotation procedure
- T6: Validating results
- T7: Delivering and maintaining the product

Questions related to annotation

1. Preparation

- Choosing the corpus — which corpus? What are the political and social ramifications?
- How to achieve balance, representativeness, and timeliness? What does it even mean?

2. ‘Instantiating’ the theory (the annotation scheme)

- Creating the annotation choices — how to remain faithful to the theory? Goal?
- Writing the manual: this is non-trivial
- Testing for stability

3. Interface design

- Building the interfaces. How to ensure speed and avoid bias?

Questions related to annotation

4. The annotators

- Choosing the annotators — what background? How many?
- How to avoid overtraining? And undertraining? How to even know?

5. Annotation procedure

- How to design the exact procedure? How to avoid biasing annotators?
- Reconciliation and adjudication processes among annotators

6. Validation

- Measuring inter-annotator agreement — which measures?
- What feedback to step 2? What if the theory (or its instantiation) ‘adjusts’?

7. Delivery

- Wrapping the result — in what form?
- Licensing, maintenance, and distribution

Choosing the corpus

T1. Prep: Choosing the corpus

- Corpus collections are worth their weight in gold
 - Should be unencumbered by copyright
 - Should be available to whole community
- Value
 - Easy-to-procure training material for algorithm development
 - Standardized results for comparison/evaluation
- Choose carefully—the future will build on your work!
 - When to re-use something? Does it really fit with your needs?
- Important sources of raw and processed text and speech
 - The Web, the BnF, etc. Lots of institutions have digitization programs!

T1. Prep: Choosing the corpus

- Technical issues: Balance, representativeness, and timeliness
 - When is a corpus representative? Do we still care about this? (“There is no data like more data”, Mercer)
 - How to balance genre, era, domain?
 - Effect of (expected) usage of corpus
 - Experts: corpus linguists or domain specialists
- Social, political, funding issues
 - How do you ensure agreement / complementarity with others?
 - How do you choose which phenomena to annotate? Need high payoff...
 - How much time and resource do you have?

Initial BookNLP-fr Corpus

Corpus

Date	Author	Title
1830	Honoré de Balzac	Sarrasine
1836	Théophile Gautier	La morte amoureuse
1841	George Sand	Pauline
1856	Victor Cousin	Madame de Hautefort
1863	Théophile Gautier	Le capitaine Fracasse
1873	Émile Zola	Le ventre de Paris
1881	Gustave Flaubert	Bouvard et Pécuchet
1882-1883	Guy de Maupassant	Mademoiselle Fifi, nouveaux contes (1)
1882-1883	Guy de Maupassant	Mademoiselle Fifi, nouveaux contes (2)
1882-1883	Guy de Maupassant	Mademoiselle Fifi, nouveaux contes (3)
1901	Lucie Achard	Rosalie de Constant, sa famille et ses amis
1903	Laure Conan	Élisabeth Seton
1904-1912	Romain Rolland	Jean-Christophe (1)
1904-1912	Romain Rolland	Jean-Christophe (2)
1917	Adèle Bourgeois	Némoville
1923	Raymond Radiguet	Le diable au corps
1926	Marguerite Audoux	De la ville au moulin
1937	Marguerite Audoux	Douce Lumière

Defining the Annotation Scheme

T2: Defining the Annotation Scheme

- Most complex question: What to annotate?
 - Goal: practical task (like BookNLP), theory building (linguistics), or both?
 - Task/theory provides annotation categories/choices
 - Problem: tradeoff between desired detail/sophistication of desired categories and practical attainability of trustworthy annotation
 - General solution: simplify categories to ensure dependable results
 - Problem: How???
- How 'deeply' to instantiate theory?
 - Design annotation scheme very carefully — simple and transparent
 - ? Depends on theory — but also (yes? how much?) on corpus and annotators
 - Do tests first, to determine what is annotatable in practice

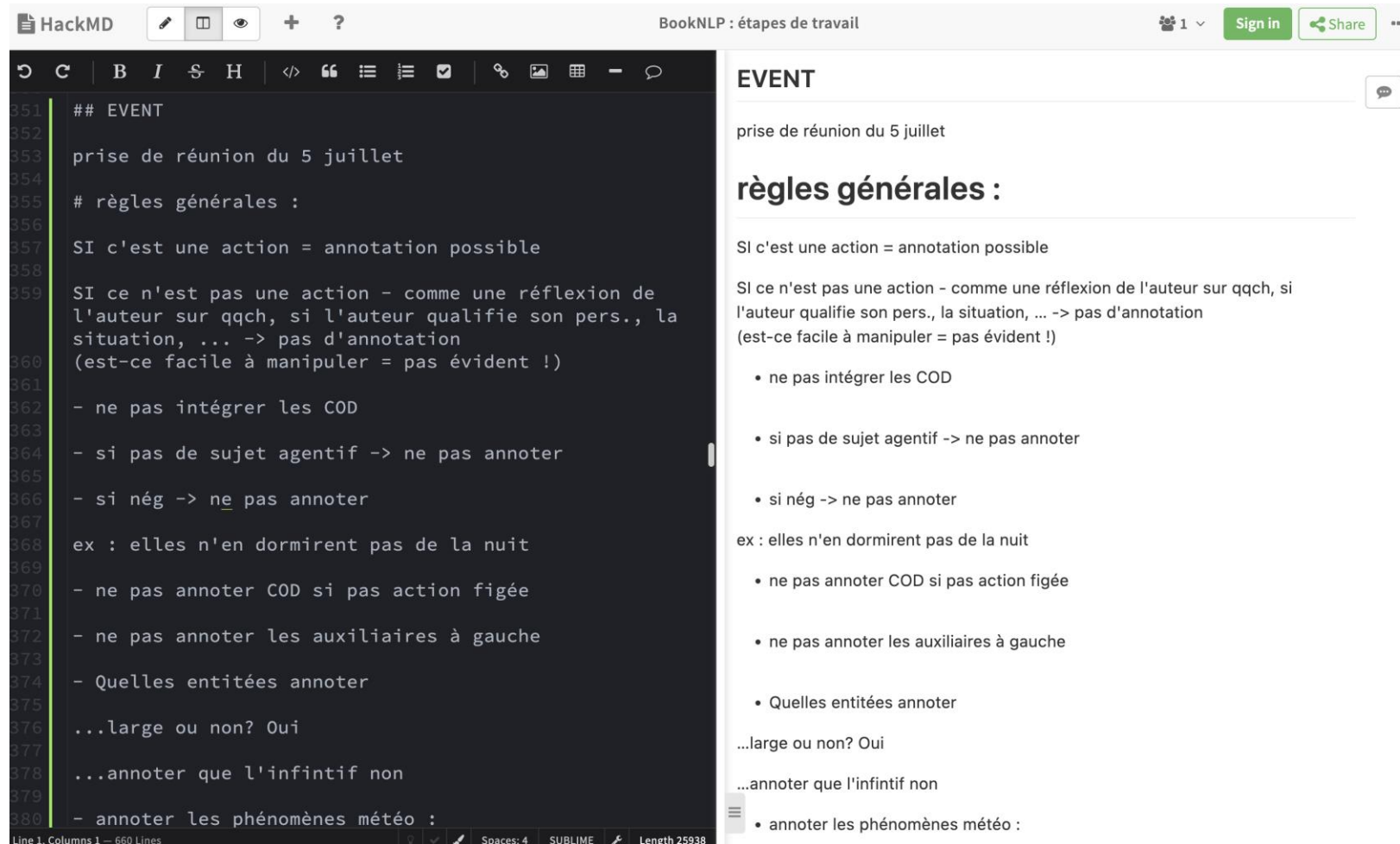
T2: Defining the Annotation Scheme

- Experts must create
 - Annotation categories
 - Annotator instruction (coding) manual — very important
 - Experts to build the manual: theoreticians? Or exactly NOT the theoreticians?
- Both must be tested! — Don't 'freeze' the manual too soon
 - Experts annotate a sample set; measure agreements
 - Annotators keep annotating a sample set until stability is achieved

T2. Defining the Annotation Scheme

- Issues
 - How to identify the right categories to annotate?
 - How easy are they to annotate? What are the borderline cases?
- Likely problems:
 - Categories not exhaustive over phenomena
 - Categories difficult to define / unclear (due to intrinsic ambiguity, or because you rely too much on background knowledge?)
- What you can do
 - Work in close cycle with annotators, and see week by week what they do
 - Hold weekly discussions with all the annotators
 - Measure the annotator agreement and disagreement (see below)
 - Modify your categories as needed—be led by what is practical
 - Create and constantly update the Annotator Handbook

BookNLP-fr Annotation Guide (with HackMD)



The screenshot displays the HackMD web interface. The top bar shows the document title "BookNLP : étapes de travail" and user options like "Sign in" and "Share". The left sidebar contains a list of documents, with "BookNLP : étapes de travail" selected. The main editor area shows a markdown document with the following content:

```
351 ## EVENT
352
353 prise de réunion du 5 juillet
354
355 # règles générales :
356
357 SI c'est une action = annotation possible
358
359 SI ce n'est pas une action - comme une réflexion de
360 l'auteur sur qqch, si l'auteur qualifie son pers., la
361 situation, ... -> pas d'annotation
362 (est-ce facile à manipuler = pas évident !)
363
364 - ne pas intégrer les COD
365
366 - si pas de sujet agentif -> ne pas annoter
367
368 - si nég -> ne pas annoter
369
370 ex : elles n'en dormirent pas de la nuit
371
372 - ne pas annoter COD si pas action figée
373
374 - ne pas annoter les auxiliaires à gauche
375
376 - Quelles entités annoter
377
378 ...large ou non? Oui
379
380 ...annoter que l'infinitif non
381
382 - annoter les phénomènes météo :
```

The right sidebar shows the rendered HTML version of the document, with the following content:

EVENT

prise de réunion du 5 juillet

règles générales :

SI c'est une action = annotation possible

SI ce n'est pas une action - comme une réflexion de l'auteur sur qqch, si l'auteur qualifie son pers., la situation, ... -> pas d'annotation (est-ce facile à manipuler = pas évident !)

- ne pas intégrer les COD
- si pas de sujet agentif -> ne pas annoter
- si nég -> ne pas annoter

ex : elles n'en dormirent pas de la nuit

- ne pas annoter COD si pas action figée
- ne pas annoter les auxiliaires à gauche
- Quelles entités annoter

...large ou non? Oui

...annoter que l'infinitif non

- annoter les phénomènes météo :

Traditional annotation issues

- Frequent borderline cases
 - Pers : Dieu, other non human characters (Zeus, animals actively speaking...) ; « un nouveau visage apparu en ville », « la foule », « la moitié de la ville » (collective nouns, fuzzy sets), « on »
 - GPE / Loc / Fac : la lande (la Lande), la route de Bressuire
 - With category Fac, what is the limit? A room, a part of a room? (but hero hiding in a closet). (cf. Bamman: everything can potentially be a location)
 - VEH : animals?
- What about robustness?

— C'est toi, la *Goualeuse* (1) — dit l'homme en blouse — tu vas me payer l'eau d'aff (2), ou je te fais danser sans violons !

— Jen'ai pas d'argent — répondit la femme en tremblant; car cet homme inspirait une grande terreur dans le quartier.

— Si ta *filoche* est à jeun (3), l'ogresse du tapis-franc te fera crédit sur ta bonne mine.

— Mon Dieu..... je lui dois déjà le loyer des vêtements que je porte.....

— Ah ! tu raisones ? — s'écria le Chourineur; et il donna dans l'ombre et au hasard un si violent coup de poing à cette malheureuse, qu'elle poussa un cri de douleur aigu.

The annotation interface

T3. The annotation interface

- How to design adequate interfaces? Re-use, avoid costly development!
- Maximize speed!
 - Create very simple tasks—but how simple? Boredom factor, but simple task means less to annotate before you have enough
 - Don't use the mouse
 - Customize the interface for each annotation project?
- Don't bias annotators (avoid priming!)
 - Beware of order of choice options
 - Is it ok to present together a whole series of choices with expected identical annotation? — annotate en bloc?
 - Do you show the annotator how 'well' he/she is doing? Why not?
- Experts: Psych experimenters, interface design specialists

Corpus Visualisation with Sacr (Oberlé, 2018)

Balthazar Balthazar , Balthazar le cheval de MmeFrancois madame François , Balthazar une bête trop grasse , tenait la tête de la file. Balthazar Il marchait, dormant à demi, dodelinant des oreilles, lorsque, à Longchamp la hauteur de la rue de Longchamp , un sursaut de peur Balthazar le planta net sur Balthazar ses quatre pieds. Les autres bêtes vinrent donner de la tête contre le cul T10 des voitures , et la file s'arrêta, avec la secousse des ferrailles, au milieu des jurements T64 des charretiers réveillés .

MmeFrancois Madame François , adossée à une planchette contre MmeFrancois ses légumes, regardait, ne voyait rien, dans la maigre lueur jetée à gauche par la petite lanterne carrée, qui n'éclairait guère qu'un des flancs luisants de Balthazar Balthazar .

Annotators

T4: Annotators

- How to choose annotators?
 - Annotator backgrounds — should they be experts, or precisely not?
 - Biases, preferences, etc.
- Who should train the annotators? Who is the most impartial?
 - Domain expert/people involved in the project
- When to train?
 - Need training session(s) before starting
 - Extremely helpful to continue weekly general discussions:
 - Identify and address hard problems
 - Expand the annotation Handbook
- BUT need to go back (re-annotate) to avoid ‘annotation drift’

T4. How much to train annotators?

- Undertrain: Instructions are too vague or insufficient. Result: annotators create their own ‘patterns of thought’ and diverge from the gold standard, each in their own particular way (Bayerl 2006)
 - How to determine? Spor indistinguishable categories, collapse them, and (?) reformulate theory — is this ok?
 - Basic choice: EITHER ‘fit’ the annotation to the annotators — is this ok? OR train annotators more — is this ok?
- Overtrain: Instructions are so exhaustive that there is no room for thought or interpretation (annotators follow a ‘table lookup’ procedure)
 - How to determine: is task simply easy, or are annotators overtrained?
 - What’s really wrong with overtraining? No predictive power...

(Fake) BookNLP pictures



The Annotation procedure

T5. Annotation procedure

- Overall approach — Shulman's rule: do the easy annotations first, so you've seen the data when you get to the harder cases
- The '85% clear cases' rule (Wiebe)
 - Ask the annotators also to mark their level of certainty
 - There should be a lot of agreement at high certainty — the clear cases
- Hypothesis (Rosé): for up to 50% incorrect instances, it pays to show the annotator possibly buggy annotations and have them correct them (compared to having them annotate anew)

T5. Annotation procedure

- Reconciliation
 - Allow annotators to discuss problematic cases, then continue — can greatly improve agreement but at the cost of drift / overtraining
- Backing off: In cases of disagreement, what do you do?
 - (1) make option granularity coarser; (2) allow multiple options; (3) increase context supporting annotation; (4) annotate only major / easy cases
- Adjudication after annotation, for the remaining hard cases
 - Have an expert (or more annotators) decide in cases of residual disagreement — but how much disagreement can be tolerated before just redoing the annotation?

T5. Annotation procedure heuristics

- Tips and possible issues
 - When annotating multiple variables, annotate each variable separately, across whole corpus — speedup and local expertise ... but lose context
 - The problem of ‘annotation drift’: shuffling and redoing items
 - Annotator attention and tiredness; rotating annotators
 - Complex management framework, interfaces, etc.
- Active learning: In-line process to dynamically find problematic cases for immediate tagging (more rapidly get to the ‘end point’), and/or to pre-annotate (help the annotator under the Rosé hypothesis)
 - Benefit: speedup; danger: misleading annotators

(Fake) BookNLP pictures



Validating annotations

T6. Validating annotations

- What to evaluate?
 - Individual agreement scores between creators
 - Overall agreement averages?
- What measure(s) to use?
 - Simple agreement is biased by chance agreement — however, this may be fine, if all you care about is a system that mirrors human behavior
 - Kappa is better for testing inter-annotator agreement. But it is not sufficient — cannot handle multiple correct choices, and works only pairwise
 - Krippendorff's alpha, Kappa variations...

T6. Validating annotations

- Tolerances
 - When is the agreement no longer good enough? — why the 90% rule? (Marcus's rule: if humans get $N\%$, systems will achieve $(N-10)\%$)
- The problem of asymmetrical/unbalanced corpora
 - When you get high agreement but low Kappa — does it matter? An unbalanced corpus makes choice easy but Kappa low. Are you primarily
 - interested in annotation qua annotation, or in doing the task?
- Experts: Psych experimenters and Corpus Analysis statisticians

T6. Agreement counts: Kappa

- Simple agreement
 - A = number choices agreed / total number
- But what about random agreement? Fix using Cohen's Kappa:
 - E = expected number of choices agreed / total number
 - $\text{Kappa} = (A - E) / (1 - E)$
- Example
 - Assume 100 examples, 50 labeled A, and 50 B: $E_{\text{random}} = 0.5$
 - Then a random annotator would score 50%: $A_{\text{random}} = 0.5$
 - $\text{Kappa}_{\text{random}} = (0.5 - 0.5) / (1 - 0.5) = 0$
 - And an annotator with 70% agreement?: $A_{70} = 0.7$
 - $\text{Kappa}_{70} = (0.7 - 0.5) / (1 - 0.5) = 0.2 / 0.5 = 0.4$
 - This is much lower than 0.7, but reflects the nonrandom agreement

T6. Limitations

- Shortcomings of Kappa
 - Works only to compare 2 annotators (else use Fleiss's Kappa)
 - Doesn't apply when multiple correct choices possible
 - Penalizes when choice distribution is skewed — but if that's the nature of the data, then why penalize?

BookNLP-fr

- Kappa
 - ??
 - (we were not so rigorous ourselves... although we measured it several times)
 - Good for entities, not so good for events

BookNLP-fr result for coreference (automatic annotation)

	recall	precision	F
bcub	74.5	65.8	69.8
muc	91.3	85.4	88.3
ceaf	65.7	81.7	72.8
average	77.2	77.6	77

BookNLP-fr result for entity recognition (automatic annotation)

	recall	precision	F
FAC	71.7	70.5	71.1
GPE	52.9	68.2	59.6
LOC	63.3	49.1	55.3
ORG	14.3	100	25.0
PER	91.5	85.2	88.2
TIME	39.7	75.3	52.0
VEH	68.2	62.5	65.2

Delivery

T7: Delivery

- It's not just about annotation...
 - How do you make sure others use the corpus?
 - FAIR principles (Findable, Accessible, Interoperable and Reusable), <http://opensciencefair.eu/>
- Technical issues
 - Licensing (CC: Creative Commons)
 - Distribution
 - Support/maintenance (over years?)
 - Incorporating new annotations/updates: layering
 - Experts: Data managers

Fr-Litbank on GitHub

The screenshot shows the GitHub repository page for `lattice-8094/fr-litbank`. The repository is public and has 6 stars and 2 forks. The repository structure is as follows:

File/Folder	Description	Last Commit
<code>MODEL_NLP-schema</code>	Add files via upload	3 years ago
<code>brat</code>	inversion du sens des coréférences dans citations : "entit...	2 years ago
<code>doc</code>	ajout manuel annot in doc	2 years ago
<code>modeling</code>	Eviter que output_dir et data_dir soient le même	2 years ago
<code>sacr</code>	suppr les anciennes annot	3 weeks ago
<code>src</code>	maj files - error espace	last year
<code>urs-xml</code>	maj des annotations	3 years ago
<code>urs</code>	maj des annotations	3 years ago
<code>xml</code>		
<code>.gitignore</code>		

The repository is licensed under a [CC BY-SA 2.0 FR](https://creativecommons.org/licenses/by-sa/2.0/fr/) license.

fr-litBank is licensed under a [Attribution-ShareAlike 2.0 France \(CC BY-SA 2.0 FR\)](https://creativecommons.org/licenses/by-sa/2.0/fr/).

Conclusion

In conclusion...

- Annotation is both
 - A mechanism for providing new training material for machine
 - A mechanism for theory formation and validation — in addition to domain specialists, annotation can involve linguists, philosophers of language, etc. in a new paradigm

Thank you for your attention!