Annotation collaborative de corpus : limiter le coût de l'annotation

Karën Fort

karen.fort@sorbonne-universite.fr / https://members.loria.fr/KFort/

LETTRES SORBONNE UNIVERSITÉ

Solutions

Solutions

- Outils d'annotation
- Propagation d'étiquettes / Pré-annotation / Apprentissage actif
- Formation / Documentation / Méthodologie
- Myriadisation (*crowdsourcing*) : `Amazon Mechanical Turk` et jeux ayant un but (`GWAP`) [prochain cours]

Pour quoi faire ?

Pourquoi utiliser des outils ?

- pour faciliter l'édition des annotations, en particulier dans le cas de relations

Pourquoi utiliser des outils ?

- pour faciliter l'édition des annotations, en particulier dans le cas de relations
- pour limiter le nombre d'items à garder en mémoire [Dandapat et al., 2009]

Pourquoi utiliser des outils ?

- pour faciliter l'édition des annotations, en particulier dans le cas de relations
- pour limiter le nombre d'items à garder en mémoire [Dandapat et al., 2009]
- pour contraindre l'annotation, et limiter ainsi les erreurs [de la Clergerie, 2008, Mikulová and Štěpánek, 2009]

Pourquoi utiliser des outils ?

- pour faciliter l'édition des annotations, en particulier dans le cas de relations
- pour limiter le nombre d'items à garder en mémoire [Dandapat et al., 2009]
- pour contraindre l'annotation, et limiter ainsi les erreurs [de la Clergerie, 2008, Mikulová and Štěpánek, 2009]
- pour cacher une couche lorsqu'on en annote une autre [Widlöcher and Mathet, 2009]

Pourquoi utiliser des outils ?

- pour faciliter l'édition des annotations, en particulier dans le cas de relations
- pour limiter le nombre d'items à garder en mémoire [Dandapat et al., 2009]
- pour contraindre l'annotation, et limiter ainsi les erreurs [de la Clergerie, 2008, Mikulová and Štěpánek, 2009]
- pour cacher une couche lorsqu'on en annote une autre [Widlöcher and Mathet, 2009]
- pour faciliter l'accès au contexte, même large [Widlöcher and Mathet, 2009]

Pourquoi utiliser des outils ?

- pour faciliter l'édition des annotations, en particulier dans le cas de relations
- pour limiter le nombre d'items à garder en mémoire [Dandapat et al., 2009]
- pour contraindre l'annotation, et limiter ainsi les erreurs [de la Clergerie, 2008, Mikulová and Štěpánek, 2009]
- pour cacher une couche lorsqu'on en annote une autre [Widlöcher and Mathet, 2009]
- pour faciliter l'accès au contexte, même large [Widlöcher and Mathet, 2009]
- pour garder une trace des discussions entre annotateurs [Lortal et al., 2006] ou des erreurs commises et de leur correction [de la Clergerie, 2008]

Quelques outils existants

+/- WebAnno, Glozz, GATE, Knowtator, Callisto, etc. :
voir https://annotationsaurus.herokuapp.com/

++ gain en temps et en qualité

⇒ (trop) nombreux outils, développés pour démontrer l'intérêt d'un schéma d'annotation ou pour une campagne spécifique, pas pour les annotateurs !

Propagation d'étiquettes (*Tag Dictionaries*)

Permet de :

1. stocker les catégories associées par les annotateurs à un token
2. proposer ces catégories lorsque ce même token est de nouveau proposé à l'annotation

$\Rightarrow$ Très simple et relativement efficace (see [Carmen et al., 2010]), mais plus on annote, plus la méthode est efficace.

Correction de pré-annotations automatiques

++ gain en temps et en qualité significatif, au moins pour l'étiquetage morpho-syntaxique et syntaxique (`Penn Treebank` [Marcus et al., 1993], étiquetage morpho-syntaxique de l'Hindi et du Bangla [Dandapat et al., 2009], étiquetage morpho-syntaxique de l'anglais [Fort and Sagot, 2010])

- biais pas toujours pris en compte : est-ce la même chose de pré-annoter des entités nommées et du renommage de gènes ?
- également consommateur de temps si le système est trop mauvais (à définir)

Cas particulier : l'apprentissage actif (*Active Learning*)

- toutes les annotations ne sont pas nécessaires pour entraîner un outil $\Rightarrow$ détecter les annotations qui sont vraiment utiles pour améliorer les résultats finaux
- pré-annoter un corpus automatiquement, puis demander aux annotateurs de corriger l'annotation, puis ré-entraîner l'outil et déterminer, grâce aux scores, ce qu'il faut revoir, etc.

$\Rightarrow$ itératif

\+ permet de gagner du temps

\- mais consommateur de temps si le système produit des résultats de trop mauvaise qualité (à définir)

- sur le projet Ritel (dialogue oral humain-machine) : plus de 30 % d'erreurs, les transcripteurs travaillaient plus vite en partant de rien plutôt qu'en corrigeant la transcription

Question autour de la pré-annotation

- soit les humains se concentrent sur ce qui a été pré-annoté et corrigent les pré-annotations, sans voir ce qui manque
- soit ils se concentrent sur ce qui manque et ne corrigent pas la pré-annotation.

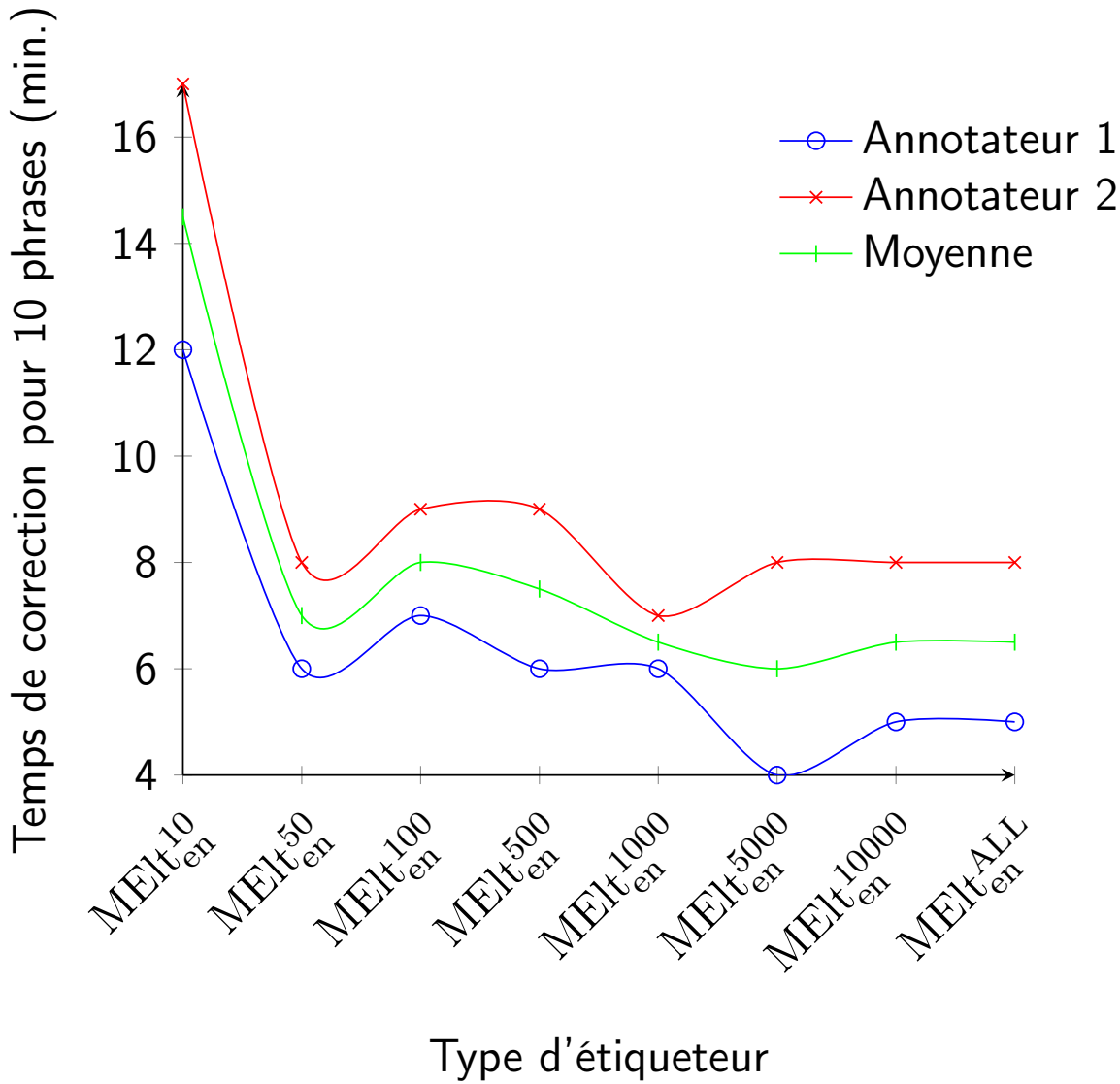
- impossible pour certains type d'annotations du fait du manque de systèmes de qualité (comme les étiqueteurs de chaînes anaphoriques)

Impact de la pré-annotation [Fort and Sagot, 2010]

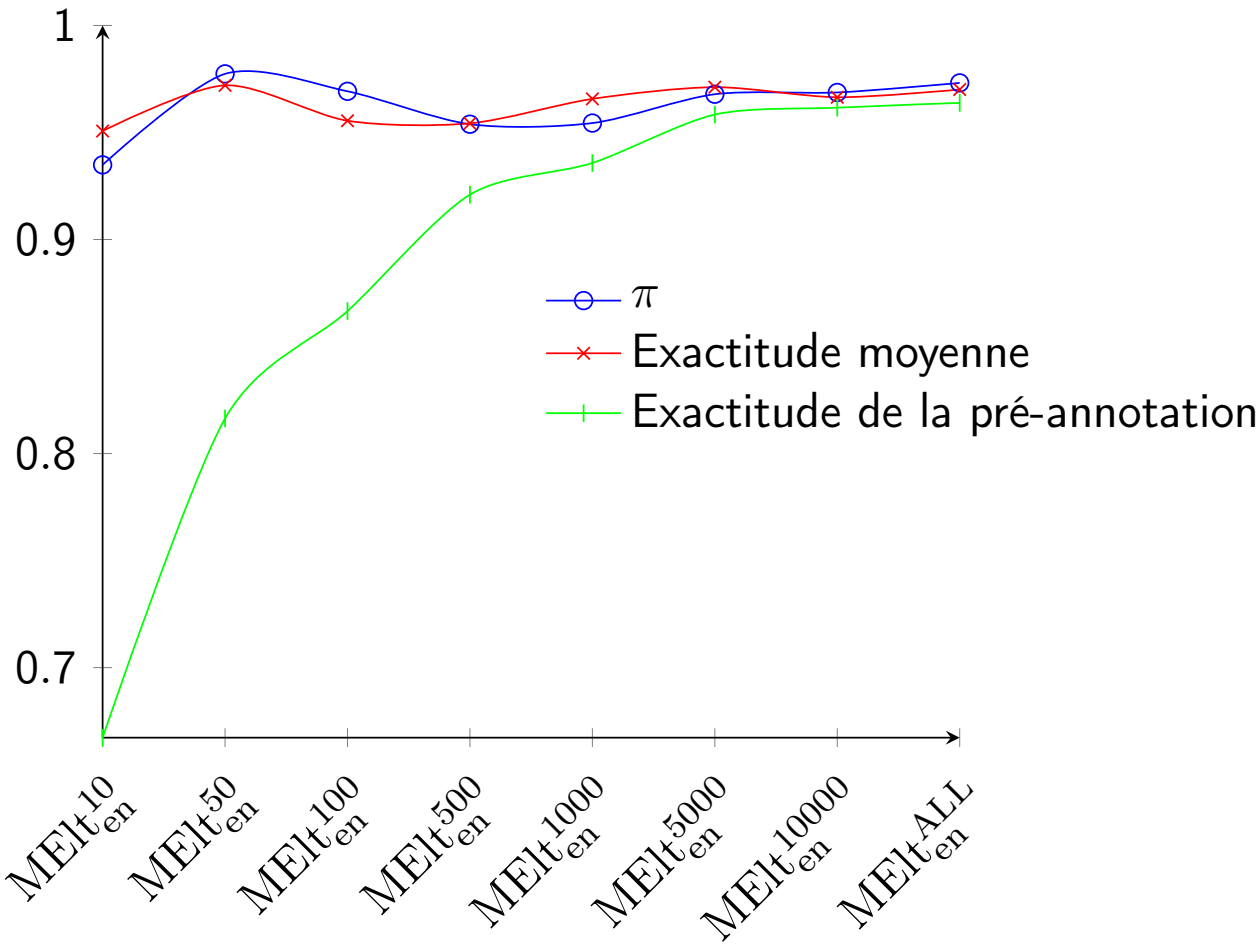
- gain en temps et en qualité (accord inter-annotateurs et exactitude)
- influence de différents niveaux de qualité de la pré-annotation
- biais introduit par la pré-annotation

... tout en limitant les **effets de la courbe d'apprentissage**

Temps de correction

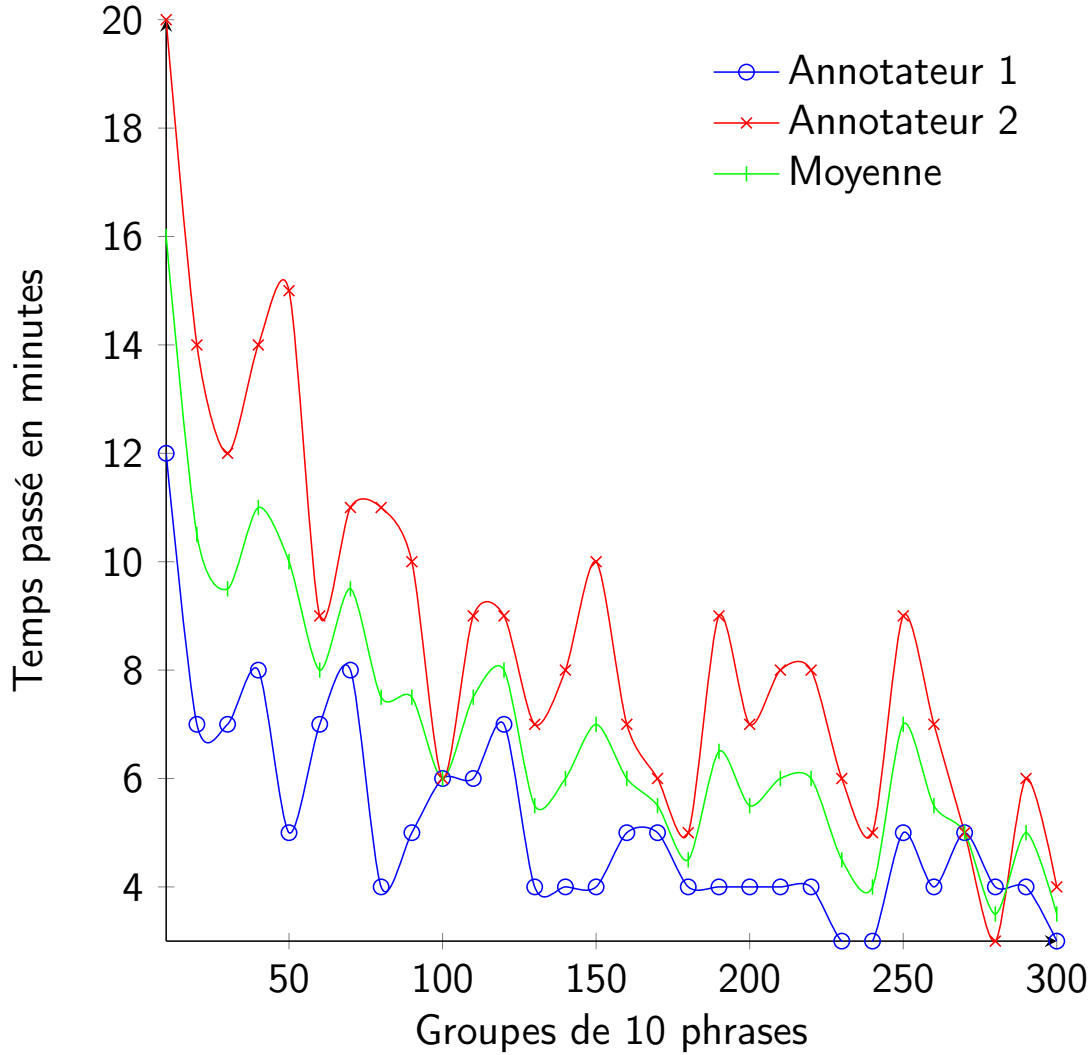


Qualité de la correction



Taille du corpus d'entraînement pour l'étiqueteur

Courbe d'apprentissage : annotation morpho-syntaxique du *Penn Treebank* [Fort and Sagot, 2010]



Les 7 maximes de Leech [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Note : peut être difficile après normalisation ("l'arbre" $\rightarrow$ "le arbre", etc.)

Les 7 maximes de Leech [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Note : peut être difficile après normalisation ("l'arbre" $\rightarrow$ "le arbre", etc.)
2. Annotations should be extractable from the text

Les 7 maximes de Leech [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Note : peut être difficile après normalisation ("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)

Les 7 maximes de Leech [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Note : peut être difficile après normalisation ("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...)

Les 7 maximes de Leech [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Note : peut être difficile après normalisation ("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)

Les 7 maximes de Leech [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Note : peut être difficile après normalisation ("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

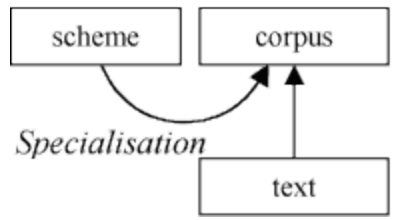
Les 7 maximes de Leech [Leech, 1993]

1. It should always be possible to come back to initial data (example BC). Note : peut être difficile après normalisation ("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)

Différents points de vue

"you only get out what you put in" [Wallis, 2007]

Approche orientée modèle



La connaissance est dans le schéma d'annotation $\Rightarrow$ le corpus est secondaire

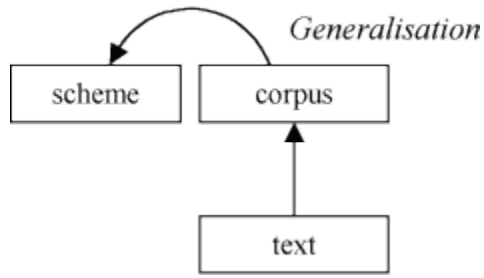
Tout est dans l'annotation !

Approche orientée modèle

- linguistique de corpus théorique
- les problèmes qui apparaissent pendant l'annotation sont :
  - résolus en modifiant l'algorithme ou
  - ignorés et considérés comme du bruit (performance)

→ en TAL ?

Approche orientée données



La connaissance est dans le texte $\Rightarrow$ le corpus est primordial
[Sinclair]

Approche orientée données

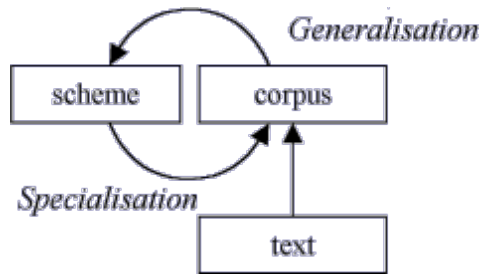
- linguistique de corpus orientée données
- "*those who select facts from theory are ignoring linguistic evidence*"
- décrire les productions linguistiques réelles et les choix faits par les locuteurs (ne pas les considérer comme de la simple 'performance')
- l'annotation est secondaire, si elle a un statut (!)
- annoter ou corriger l'analyse n'a pas de sens (!)

→ étude des collocations, concordances, patrons lexicaux

Approche orientée données

- mais succès de l'étiquetage morpho-syntaxique !
- Aujourd'hui : annotation du "minimum nécessaire"
- Combien faut-il (est-il utile d') annoter ?

Troisième voie ?



La connaissance est dans le schéma et dans le corpus

Annotation par cycles

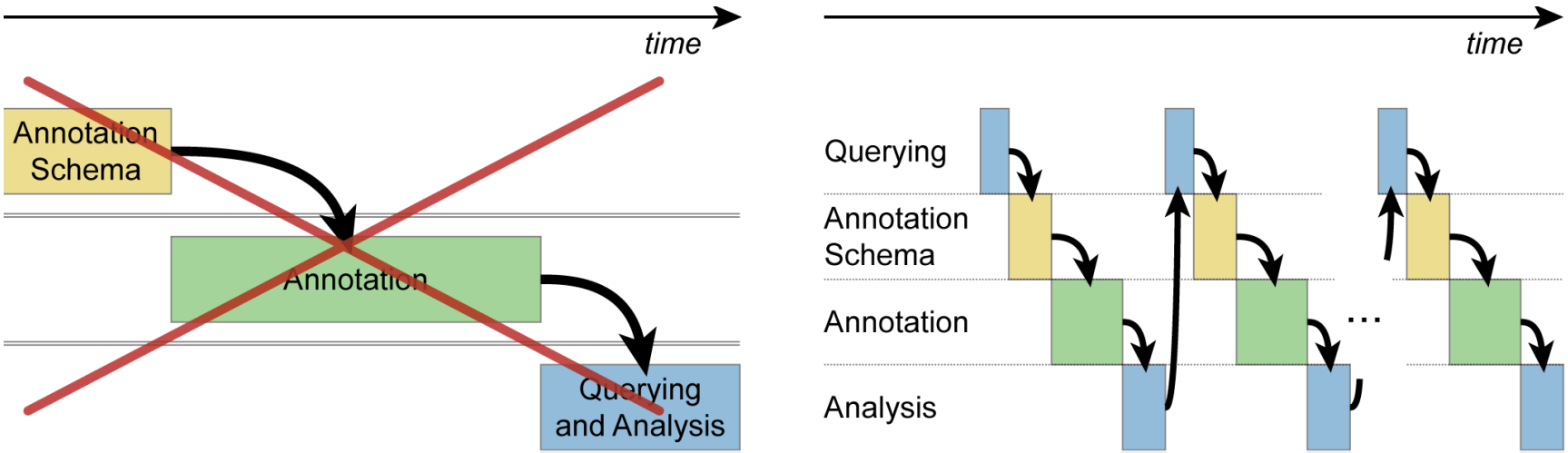
- les nouvelles observations généralisent les hypothèses
- la théorie permet d'interpréter et de classifier l'information

- cycles évolutifs : chaque cycle améliore les connaissances en raffinant et testant les théories sur de vraies données

$\Rightarrow$ une représentation du corpus plus précise est construite et un étiqueteur morpho-syntaxique (par ex) plus sophistiqué est produit

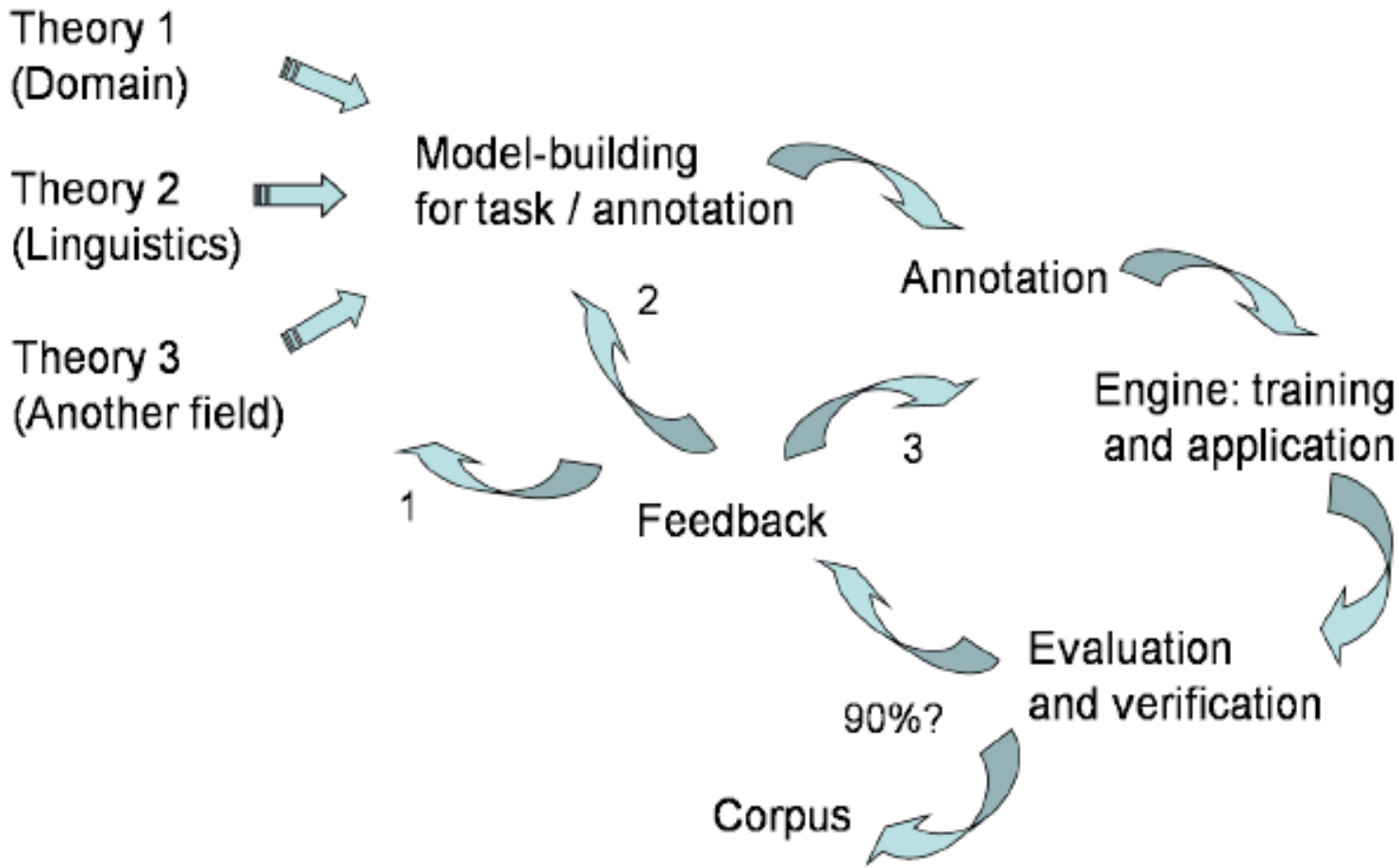
Annotation agile

intégrer l'évaluation

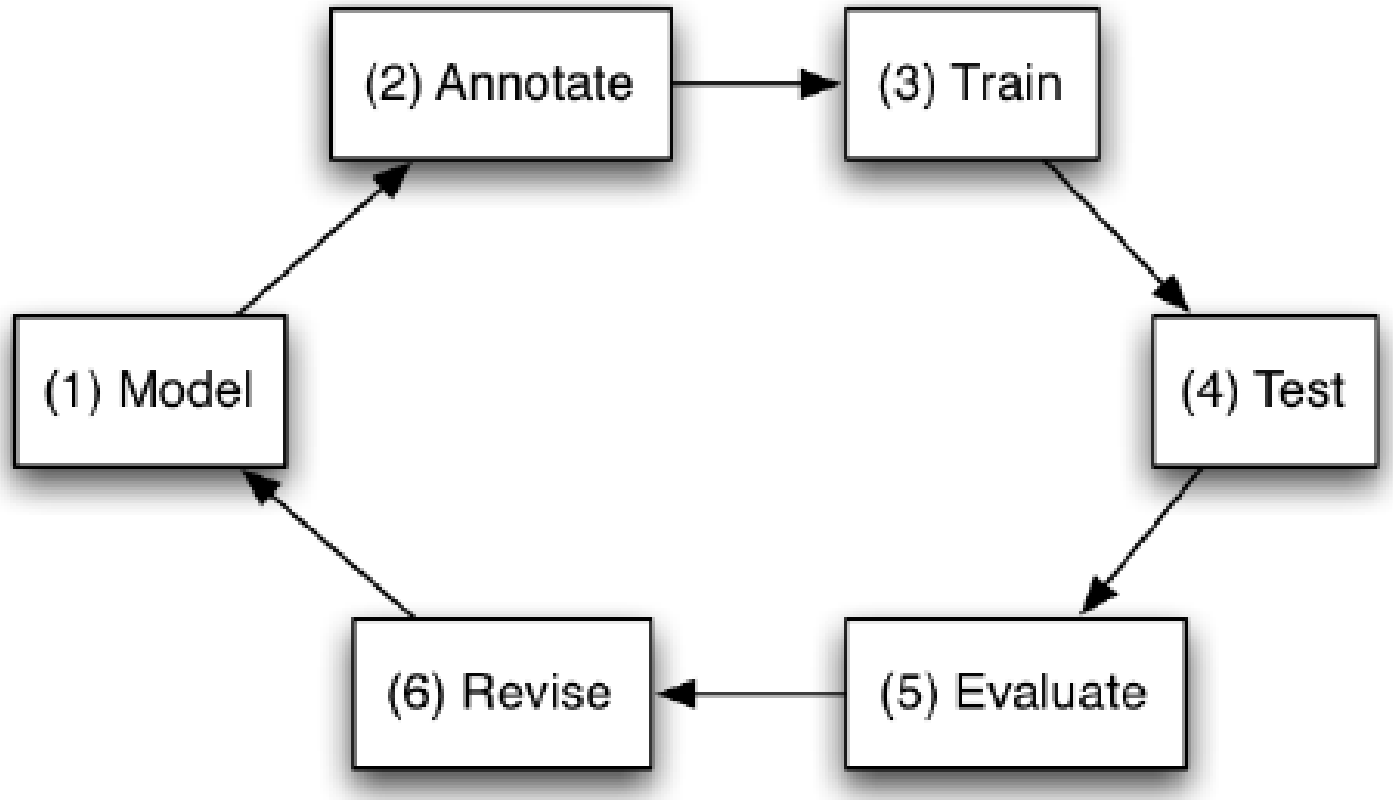


Phases de l'annotation traditionnelle (à gauche) et cycles de l'annotation agile (à droite). Reproduction de la figure 2 de [Voormann and Gut, 2008]

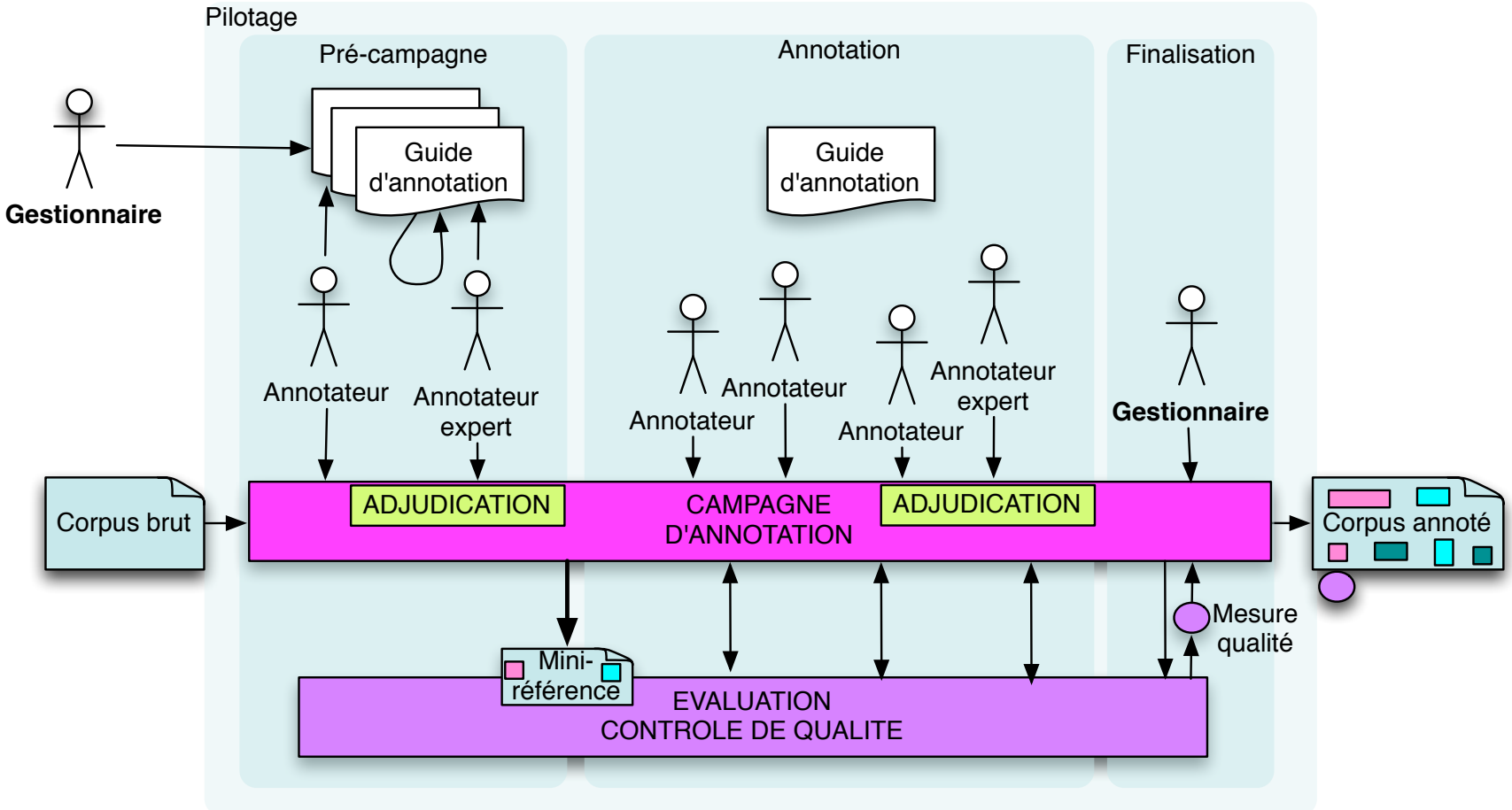
Pipeline d'annotation générique [Hovy and Lavid, 2010]



Cycle MATTER [Pustejovsky and Stubbs, 2012]



Vers un « génie annotationnel »



Méthodologie : quelques grands principes

- calculer l'accord inter-annotateurs au tout début de la campagne, puis mettre à jour le guide d'annotation [Bonneau-Maynard et al., 2005].
- calculer l'accord intra-annotateur au fur et à mesure de la campagne, pour vérifier que les annotateurs sont cohérents avec eux-mêmes [Gut and Bayerl, 2004].
- on peut aller jusqu'à faire de l'annotation agile [Voormann and Gut, 2008, Alex et al., 2010], qui implique plusieurs itérations

Formation et documentation

Une bonne formation des annotateurs est la solution la plus efficace pour gagner en temps et en qualité d'annotation [Dandapat et al., 2009].

Cela doit être associé à une documentation adaptée proposant :

- une définition claire de l'**application**
- une définition claire et détaillée des **catégories** (toujours possible ou même souhaitable ?)
- des **exemples** bien choisis
- une présentation à part des **catégories ambiguës**, comme dans la documentation du PTB (voir : ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz)

Formation et documentation

Une bonne formation des annotateurs est la solution la plus efficace pour gagner en temps et en qualité d'annotation [Dandapat et al., 2009].

Cela doit être associé à une documentation adaptée proposant :

- une définition claire de l'**application**
- une définition claire et détaillée des **catégories** (toujours possible ou même souhaitable ?)
- des **exemples** bien choisis
- une présentation à part des **catégories ambiguës**, comme dans la documentation du PTB (voir : ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz)

Ne pas oublier que les annotateurs sont au cœur de la campagne d'annotation !

- solutions
- avantages et inconvénients

Lire un article de recherche

Lire pour la semaine prochaine

[Snow et al., 2008] :
Snow, R., O'Connor, B., Jurafsky, D., and Ng., A. Y. (2008).
Cheap and fast - but is it good ? evaluating non-expert annotations for natural language tasks.
Actes de EMNLP 2008, pages 254–263.

http://web.stanford.edu/~jurafsky/amt.pdf


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.


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.


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


Dandapat, S., Biswas, P., Choudhury, M., and Bali, K. (2009).

Complex linguistic annotation - no easy way out ! a case from bangla and hindi POS labeling tasks.
In Proceedings of the third ACL Linguistic Annotation Workshop, Singapour.


de la Clergerie, E. V. (2008).
A collaborative infrastructure for handling syntactic annotations.
In Proceedings of the First International Workshop on Automated Syntactic Annotations for interoperable Language Resources, Hong-Kong, Chine.


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


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.


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).


Leech, G. (1993).
Corpus annotation schemes.
Literary and Linguistic Computing, 8(4) :275–281.


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


Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993).

Building a large annotated corpus of English : The Penn Treebank.
Computational Linguistics, 19(2) :313–330.


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


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


Snow, R., O'Connor, B., Jurafsky, D., and Ng., A. Y. (2008).
Cheap and fast - but is it good ? evaluating non-expert annotations for natural language tasks.
In Proceedings of EMNLP 2008, pages 254–263.


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.


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