Axe 2 : Linguistique et évaluation des systèmes de traitement automatique des langues

Karën Fort et Guillaume Wisniewski

karen.fort@sorbonne-universite.fr, guillaume.wisniewski@limsi.fr

GDR LIFT, 4 juin 2019

L'axe 2 du GDR LIFT : la quête du graal

Linguistique et TAL, en 2019

Catégorisez, catégorisez, il en restera toujours quelque chose

Des pistes pour l'axe 2

L'axe 2 du GDR LIFT : la quête du graal

Linguistique et TAL, en 2019

Catégorisez, catégorisez, il en restera toujours quelque chose

Des pistes pour l'axe 2

L'axe 2 : des objectifs affichés...

1. explorer dans quelle mesure des critères linguistiques peuvent être utilisés pour évaluer, de façon automatique, la qualité linguistique des résultats produits par les systèmes de TAL
2. construire, à partir des recherches conduites en linguistique formelle, des batteries de tests linguistiques qui permettent d'évaluer les systèmes de TAL

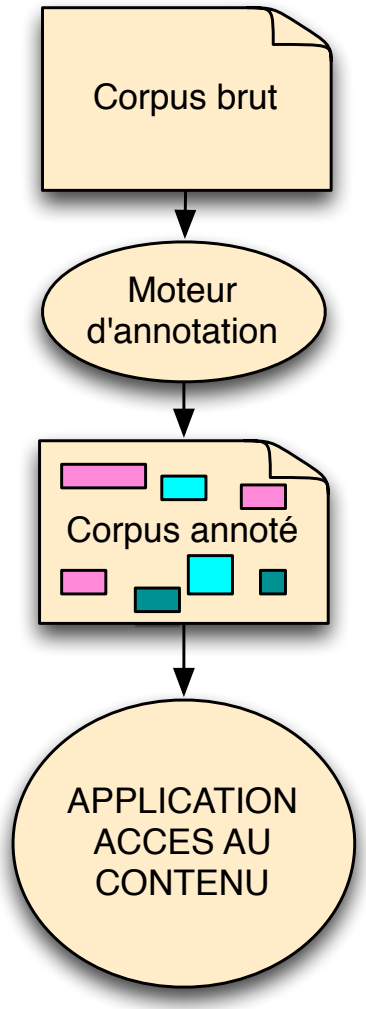
. . . et une volonté partagée

remettre la linguistique au cœur du TAL

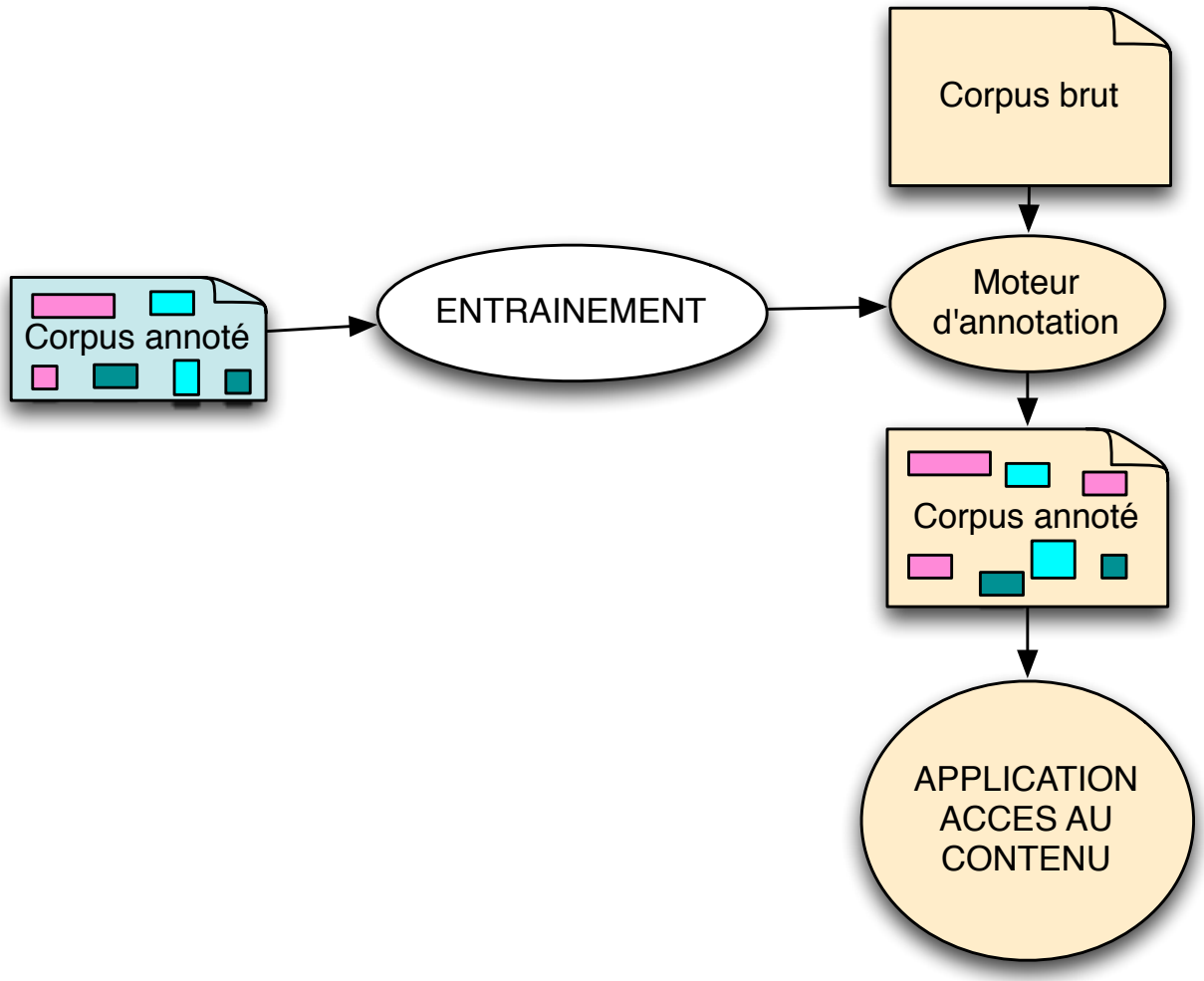
Collecte, annotation et analyse de données textuelles et sonores pour l'analyse linguistique

- Guillaume Wisniewski
- en cours de validation

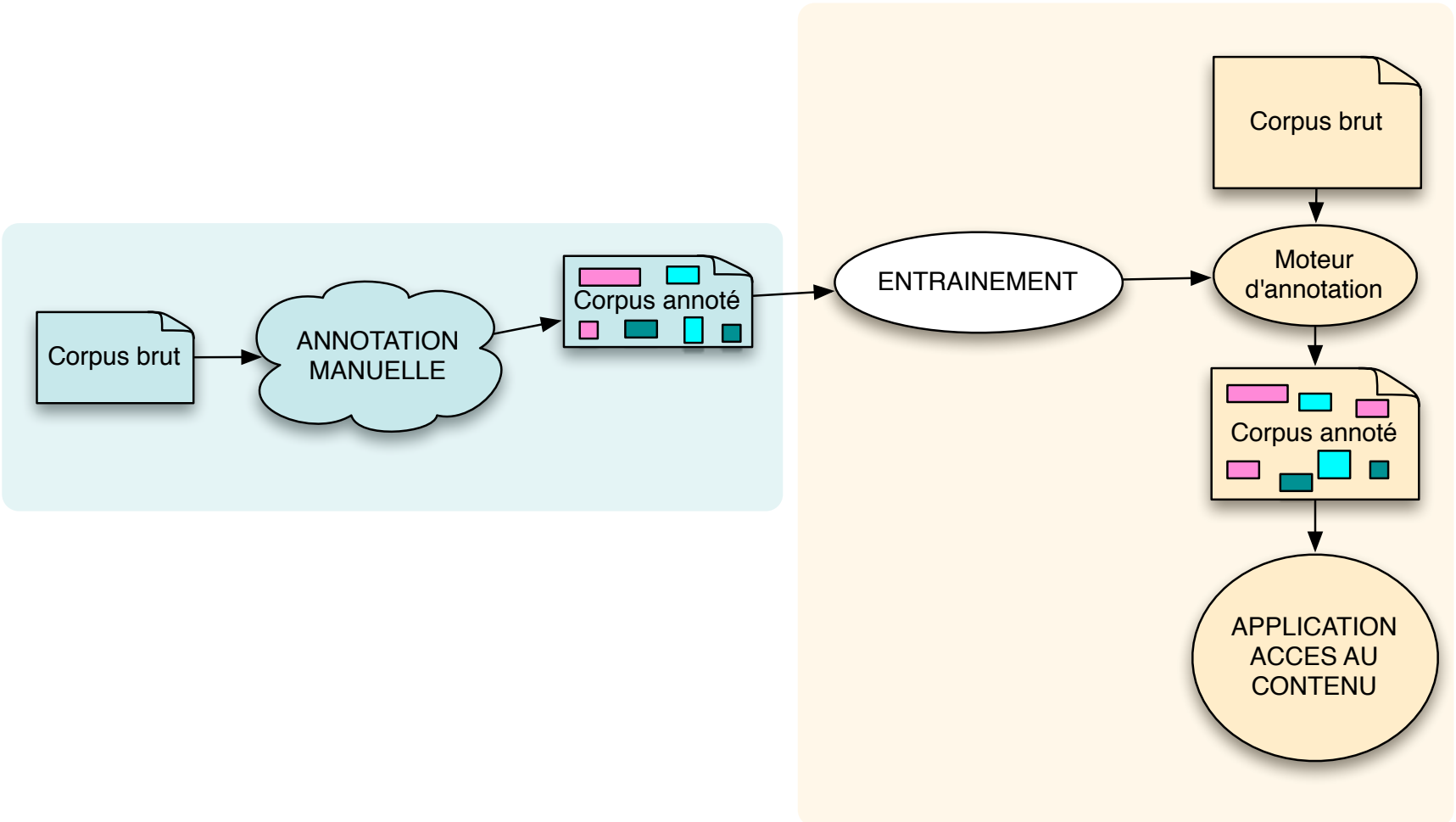
Approches supervisées et linguistique



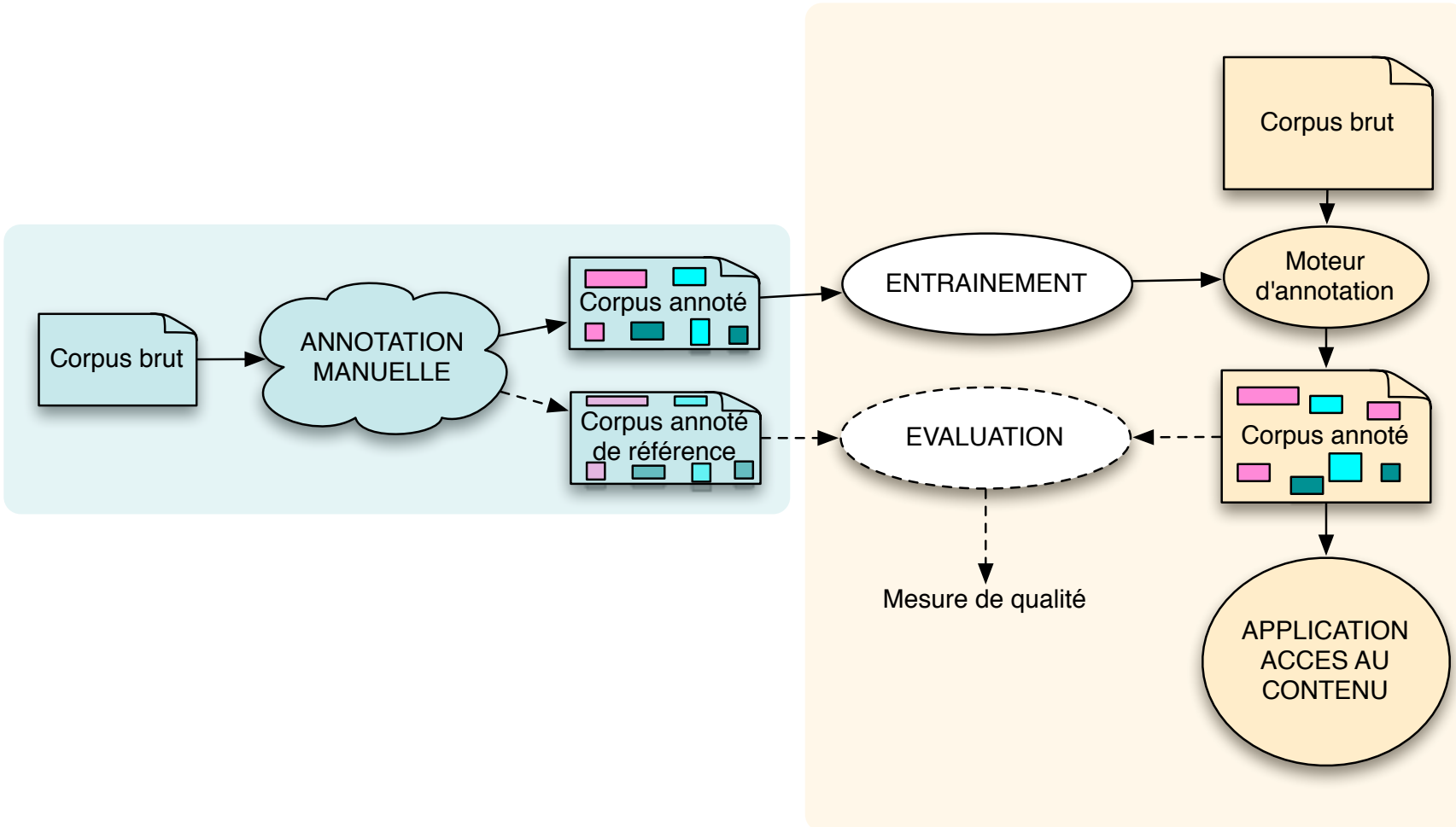
Approches supervisées et linguistique



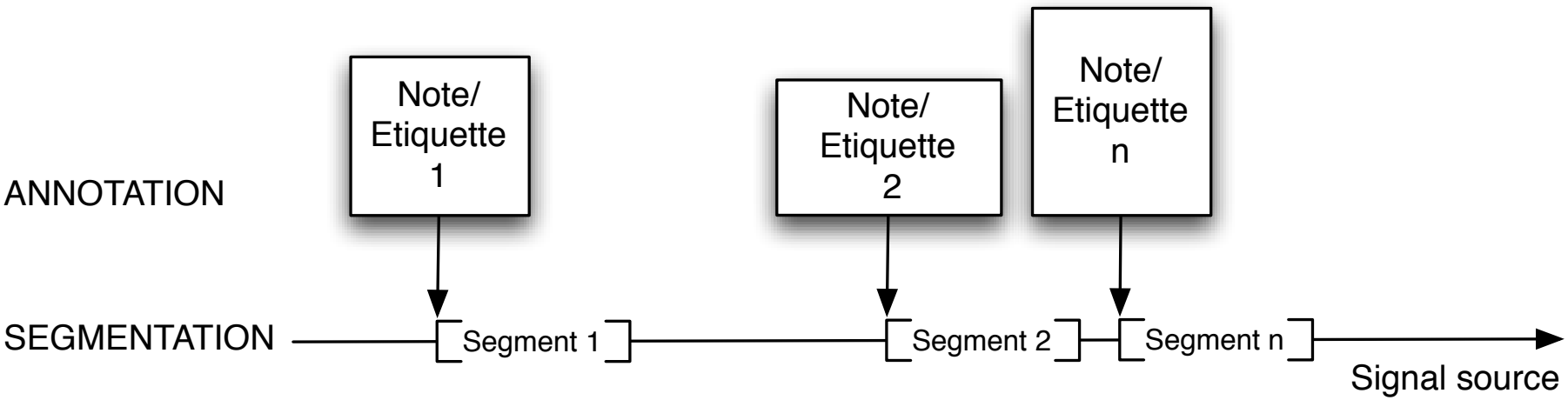
Approches supervisées et linguistique



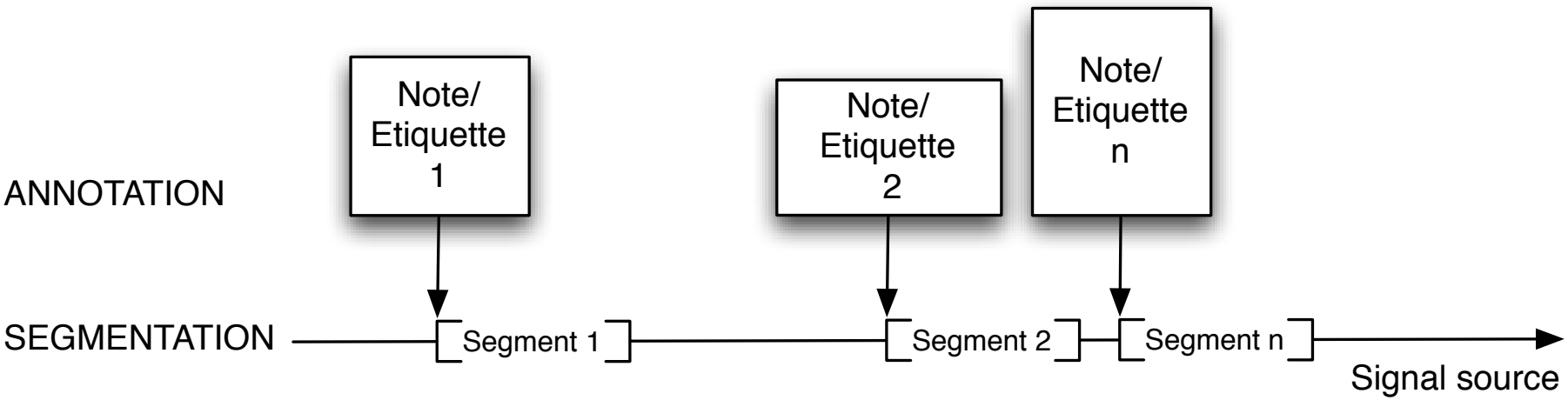
Approches supervisées et linguistique



Annotation


ANNOTATION
Note/
Etiquette
1
Note/
Etiquette
2
Note/
Etiquette
n
SEGMENTATION
Segment 1
Segment 2
Segment n
Signal source

Annotation



Ajout d'informations interprétatives [Leech, 1997, Habert, 2005]

Durée de vie des corpus annotés

Penn Treebank [Marcus et al., 1993] :

- créé au début des années 90
- encore utilisé dans les années 2010

vs tagger PARTS [Church, 1988], qui n'est plus du tout utilisé

$\rightarrow$ évolution rapide des outils

$\Rightarrow$ l'annotation ne doit **pas** dépendre d'eux

Le consensus, au cœur de l'annotation

Il faut ≪convenir pour mesurer ≫[Desrosières, 2008]

L'annotation est de l'ordre de la quantification

Mesurer *vs* quantifier [Desrosières, 2008] :

- **mesurer** : implique une forme mesurable (par ex. la hauteur du Mont Blanc)
- **quantifier** : suppose des conventions d'équivalences préalables

Outiller le consensus :

- guide d'annotation (12 p. pour le football)
- réunions avec les annotateurs et le gestionnaire de la campagne
- évaluer le consensus (la cohérence)

Jugement d'acceptabilité *vs* annotation

tel Monsieur Jourdain...

(a') *Certains libraires vendent ces livres*
(b') *Ces livres, certains libraires les vendent*
(a") *?Des libraires vendent ces livres*
(b") **Ces livres, certains libraires vendent*

[Guentchéva and Desclés, 1991]

Annotation insérée en début de phrase, 3 catégories possibles :

- acceptable (aucune note)
- non acceptable : *
- incertain : ?

Jugement d'acceptabilité *vs* annotation

tel Monsieur Jourdain...

(a') *Certains libraires vendent ces livres*
(b') *Ces livres, certains libraires les vendent*
(a") *?Des libraires vendent ces livres*
(b") **Ces livres, certains libraires vendent*

[Guentchéva and Desclés, 1991]

Obtention d'un consensus d'acceptabilité [Habert, 2008] :

- jugement éduqué, informé, soumis à un apprentissage
- suppose un travail collectif

Jugement d'acceptabilité *vs* annotation

cependant...

En France [Habert, 2008] :

- pas de guide d'acceptabilité : pas de "trace" globale des acceptabilités sur tel ou tel phénomène (sauf LADL/M. Gross)
- pas de travail en largeur ou systématique (sauf LADL/M. Gross)
- travail sur des énoncés simplifiés [Milner, 1989]

- l'annotation traite d'une très (plus ?) large variété de phénomènes (cf football)

Conséquences

Question fondamentale : **les annotations sont-elles correctes ?**

- les systèmes apprennent les erreurs des annotateurs humains (bruit $\neq$ régularités dans les erreurs [Reidsma and Carletta, 2008])
- l'évaluation (des outils) peut être faussée
- les résultats d'analyse linguistique ou de systèmes symboliques peuvent être faussés et non concluant

Validité *vs* fiabilité [Artstein and Poesio, 2008]

- nous nous intéressons à la **validité** de l'annotation manuelle
  - *i.e.* si les catégories annotées sont correctes
- Mais il n'existe pas de "vérité terrain"
  - les catégories linguistiques sont déterminées par le jugement humain
  - conséquence : il est impossible de mesurer directement si une catégorie est correcte
- nous ne pouvons mesurer que la **fiabilité** de l'annotation
  - *i.e.* si les annotateurs humains prennent les mêmes décisions de manière **cohérente** $\Rightarrow$ ils ont internalisé le schéma d'annotation
  - hypothèse sous-jacente : une fiabilité élevée implique la validité de l'annotation
- coefficient d'accord (inter-annotateurs)

L'insuffisance des accords inter-annotateurs

`Polish Treebank` [Woliński et al., 2011]

- explorent certains des cas où les annotateurs sont en accord

$\rightarrow$ 20 % d'erreurs

L'axe 2 du GDR LIFT : la quête du graal

Linguistique et TAL, en 2019

Catégorisez, catégorisez, il en restera toujours quelque chose

Des pistes pour l'axe 2

(Mes) questions de catégorisation

- comment écrire un guide d'annotation :
  - efficace
  - sans biais
- comment rendre l'annotation humaine plus efficace pour les machines : *active learning*, mais du point de vue humain
- existence et évolution des catégories dans le TAL (entités nommées, POS) et en linguistique [Haspelmath, 2015]

(Vos) questions/propositions

Artstein, R. and Poesio, M. (2008).
Inter-coder agreement for computational linguistics.
Computational Linguistics, 34(4) :555–596.

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 I.
Presses de l'École des Mines de Paris.

Guentchéva, Z. and Desclés, J.-P. (1991).
Test et acceptabilité.
Histoire Épistémologie Langage, 13(2) :9–25.

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).


Habert, B. (2008).
Observer, aujourd'hui, c'est manipuler.
In François, J., editor,
Observations et manipulations en linguistique : entre concurrence et c
volume 16 of Mémoires de la Société de linguistique de Paris.
Nouvelle série, pages 33–53. Peeters, Paris, France.


Haspelmath, M. (2015).
How categorical are categories ?, chapter Defining vs diagnosing linguistic categories : a case study of clitic phenomena.

De Gruyter Mouton.


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


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.


Milner, J. (1989).
Introduction à une science du langage.
Des travaux. Editions du Seuil.


Reidsma, D. and Carletta, J. (2008).

Reliability measurement without limits.
Computational Linguistics, 34(3) :319–326.

Woliński, M., Głowińska, K., and Świdziński, M. (2011).
A preliminary version of składnica - a treebank of polish.
In Proceedings of the 5th Language and Technology Conference.