Annotation collaborative de corpus : Dimensions de complexité

Karën Fort

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

LETTRES SORBONNE UNIVERSITÉ

Annoter manuellement

En quoi est-ce facile ou difficile ? Quoi outiller ?

Parties du discours [Marcus et al., 1993] :

I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.

Annoter manuellement

En quoi est-ce facile ou difficile ? Quoi outiller ?

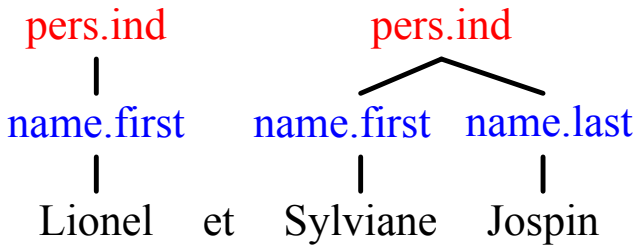
Renommage de gènes [Fort et al., 2012a] :

> *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.*

Annoter manuellement

En quoi est-ce facile ou difficile ? Quoi outiller ?

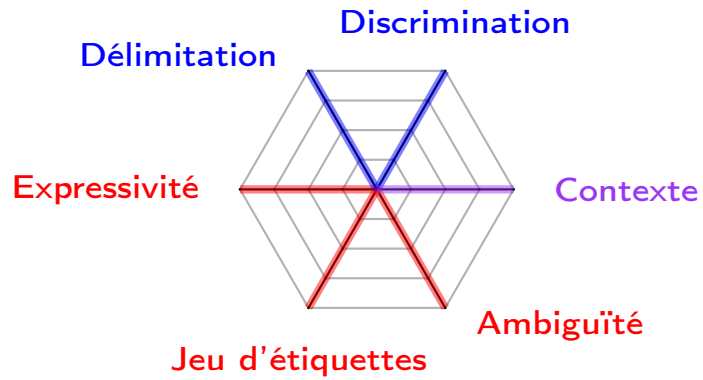
Entités nommées structurées [Grouin et al., 2011] :



Qu'est-ce qui est complexe ?

Dimensions de complexité [Fort et al., 2012b]

1. Discrimination des unités à annoter
2. Délimitation des unités à annoter
3. Expressivité du langage d'annotation
4. Dimension du jeu d'étiquettes
5. Ambiguïté
6. Contexte à prendre en compte



- Métriques associées, calculables a priori ou sur un échantillon
- Indépendantes du volume à annoter et du nombre d'annotateurs

Discrimination

Parties du discours [Marcus et al., 1993], pré-annotées :

I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.

Renommage de gènes [Fort et al., 2012a], non pré-annoté :

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.

⇒ **plus difficile** si les unités à annoter sont « noyées » au milieu des autres, en particulier si la segmentation n'est pas évidente.

Discrimination

Plus la proportion de ce qui *doit* être annoté par rapport à ce qui *pourrait* être annoté est faible, plus le poids de la discrimination est élevé :

Définition

$$Discrimination(Flux) = 1 - \frac{|Annotations(Flux)|}{\sum_{i=1}^{nivSeg} |UtésObtParDécoupage_i(Flux)|}$$

$\Rightarrow$ Nécessité d'une segmentation de référence

Parties du discours [Marcus et al., 1993] :

I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.

$$Discrimination_{PTB_{POS}} = 0$$

Renommage de gènes [Fort et al., 2012a] :

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.

$$Discrimination_{Renommage} = 0,95$$

Délimitation des frontières

Délimiter les frontières consiste à :

- **étendre** ou **rétrécir** l'unité discriminée :
  Madame Chirac $\rightarrow$ *Monsieur et Madame Chirac*

Délimitation des frontières

Délimiter les frontières consiste à :

- **étendre** ou **rétrécir** l'unité discriminée :
  Madame Chirac $\rightarrow$ *Monsieur et Madame Chirac*

- **décomposer** une unité discriminée en plusieurs éléments :
  le préfet Érignac $\rightarrow$ *le préfet Érignac*

Délimitation des frontières

Délimiter les frontières consiste à :

- **étendre** ou **rétrécir** l'unité discriminée :
  Madame Chirac $\rightarrow$ *Monsieur et Madame Chirac*
- **décomposer** une unité discriminée en plusieurs éléments :
  le préfet Érignac $\rightarrow$ *le préfet Érignac*
- ou **regrouper** plusieurs unités discriminées en une seule annotation :
  Sa Majesté
  le roi Mohamed VI $\rightarrow$ *Sa Majesté le roi Mohamed VI*

Délimitation

Définition

$$Délimitation(Flux) = min\left(\frac{Subst. + Ajouts + Suppr.}{|Annotations(Flux)|}, 1\right)$$

$$Délimitation_{Renommage} = 0$$

$$Délimitation_{EN_{TypesSoustypes}} = 1$$

Gérer la complexité de l'annotation manuelle

Quoi annoter ?

Comment annoter ?

- Expressivité du langage d'annotation
- Dimension du jeu d'étiquettes
- Degré d'ambiguïté

Le poids du contexte

Outiller à bon escient

Pour finir

Expressivité du langage d'annotation

Définition

Les degrés d'expressivité du langage d'annotation sont les suivants :

- $0,25$: langages de types
- $0,5$: langages relationnels d'arité 2
- $0,75$: langages relationnels d'arité supérieure à 2
- 1 : langages d'ordre supérieur

$$Expressivité_{Renommage} = 0,25$$

$$Expressivité_{PTB_{POS}} = 0,25$$

Gérer la complexité de l'annotation manuelle

Quoi annoter ?

Comment annoter ?

- Expressivité du langage d'annotation
- Dimension du jeu d'étiquettes
- Degré d'ambiguïté

Le poids du contexte

Outiller à bon escient

Pour finir

Dimension du jeu d'étiquettes

<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 (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>
<tr><th colspan="2">Organization</th><th colspan="2" rowspan="3">Time</th></tr>
<tr><td>org.adm (administration)</td><td>org.ent (services)</td></tr>
<tr><th colspan="2">Amount</th></tr>
<tr><td colspan="2">amount (with unit or general object), including duration</td><td>time.date.abs (absolute date), time.date.rel (relative date)</td><td>time.hour.abs (absolute hour), time.hour.rel (relative hour)</td></tr>
</table>

Types et sous-types utilisés pour l'annotation en EN structurées

Dimension du jeu d'étiquettes

<table>
<tr><td colspan="3">Person</td><td colspan="3">Function</td></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><td colspan="3">Location</td><td colspan="3">Production</td></tr>
<tr><td rowspan="4">administrative (loc.adm.town, loc.adm.reg, loc.adm.nat, loc.adm.sup)</td><td rowspan="4">physical (loc.phys.geo, loc.phys.hydro, loc.phys.astro)</td><td rowspan="4">facilities (loc.fac), oronyms (loc.oro), address (loc.add.phys, loc.add.elec)</td><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>
<tr><td colspan="3">Organization</td><td colspan="3" rowspan="3">Time</td></tr>
<tr><td colspan="2">org.adm (administration)</td><td>org.ent (services)</td></tr>
<tr><td colspan="3">Amount</td></tr>
<tr><td colspan="3">amount (with unit or general object), including duration</td><td colspan="2">time.date.abs (absolute date), time.date.rel (relative date)</td><td>time.hour.abs (absolute hour), time.hour.rel (relative hour)</td></tr>
</table>

Niveau 1 : *pers*, *func*, *loc*, *prod*, *org*, *time*, *amount* $\rightarrow$ 7 possibilités (degré de liberté = 6).

Dimension du jeu d'étiquettes

<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"></td><td>prod.doctr (doctrine)</td><td>prod.rule (law)</td><td>prod.soft (software)</td></tr>
<tr><td colspan="3"></td><td>prod.art</td><td>prod.media</td><td>prod.award</td></tr>
<tr><th colspan="3">Organization</th><th colspan="3">Time</th></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></tr>
</table>

Niveau 1 : *pers*, *func*, *loc*, *prod*, *org*, *time*, *amount* $\rightarrow$ 7 possibilités (degré de liberté = 6).
Niveau 2 : *prod.object*, *prod.serv*, *prod.fin*, *prod.soft*, *prod.doctr*, *prod.rule*, *prod.art*, *prod.media*, *prod.award* $\rightarrow$ 9 possibilités (degré de liberté = 8).

Dimension du jeu d'étiquettes

Person	
pers.ind (individual person)	*pers.coll* (group of persons)

Function	
func.ind (individual function)	*func.coll* (collectivity of functions)

Location		
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*)

Production		
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	
org.adm (administration)	*org.ent* (services)

Amount
amount (with unit or general object), including duration

Time	
time.date.abs (absolute date), *time.date.rel* (relative date)	*time.hour.abs* (absolute hour), *time.hour.rel* (relative hour)

Niveau 1 : *pers*, *func*, *loc*, *prod*, *org*, *time*, *amount* $\rightarrow$ 7 possibilités (degré de liberté = 6).
Niveau 2 : *prod.object*, *prod.serv*, *prod.fin*, *prod.soft*, *prod.doctr*, *prod.rule*, *prod.art*, *prod.media*, *prod.award* $\rightarrow$ 9 possibilités (degré de liberté = 8).
Niveau 3 : *loc.adm.town*, *loc.adm.reg*, *loc.adm.nat*, *loc.adm.sup* $\rightarrow$ 4 possibilités (degré de liberté = 3).

Dimension du jeu d'étiquettes

Degré de liberté

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

où ν_i est le degré de liberté maximal que l'annotateur a dans le choix de la i^{eme} sous-étiquette ($\nu_i = n_i - 1$).

Dimension du jeu d'étiquettes

$$Dimension(Flux) = min(\frac{\nu}{\tau}, 1)$$

où τ est le seuil à partir duquel on considère le jeu d'étiquettes comme arbitrairement grand (déterminé expérimentalement).

$$Dimension_{Renommage} = 0,04$$

$$Dimension_{EN_{TypesSoustypes}} = 0,34$$

Gérer la complexité de l'annotation manuelle

Quoi annoter ?

Comment annoter ?

- Expressivité du langage d'annotation
- Dimension du jeu d'étiquettes
- Degré d'ambiguïté

Le poids du contexte

Outiller à bon escient

Pour finir

Degré d'ambiguïté : ambiguïté résiduelle

Utiliser les traces laissées par les annotateurs :

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

Définition

$$Ambiguité_{Res}(Flux) = \frac{|Annotations_{amb}|}{|Annotations|}$$

$$AmbiguitéRes_{Renommage} = 0,02$$

→ Ne s'applique pas au Penn Treebank (pas de traces).

Degré d'ambiguïté : ambiguïté théorique

Proportion des unités à annoter qui correspond à des vocables ambigus.

Définition

$$Ambiguité_{Th}(Flux) = \frac{\sum_{voc_i=1}^{|Voc(Flux)|} (Ambig(voc_i) * freq(voc_i, Flux))}{|Unités(Flux)|}$$

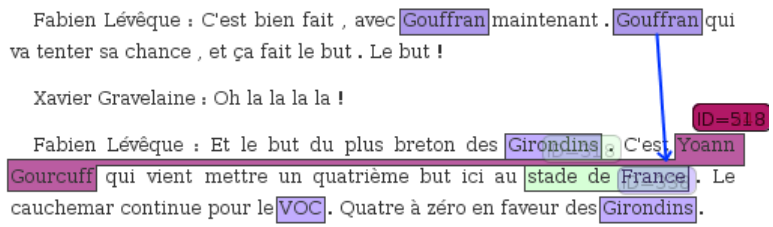
avec

$$Ambig(voc_i) = \begin{cases} 1 & \text{si} \quad |Étiquettes(voc_i)| > 1 \\ 0 & \text{sinon} \end{cases}$$

$\rightarrow$ Ne s'applique pas aux relations de renommage.

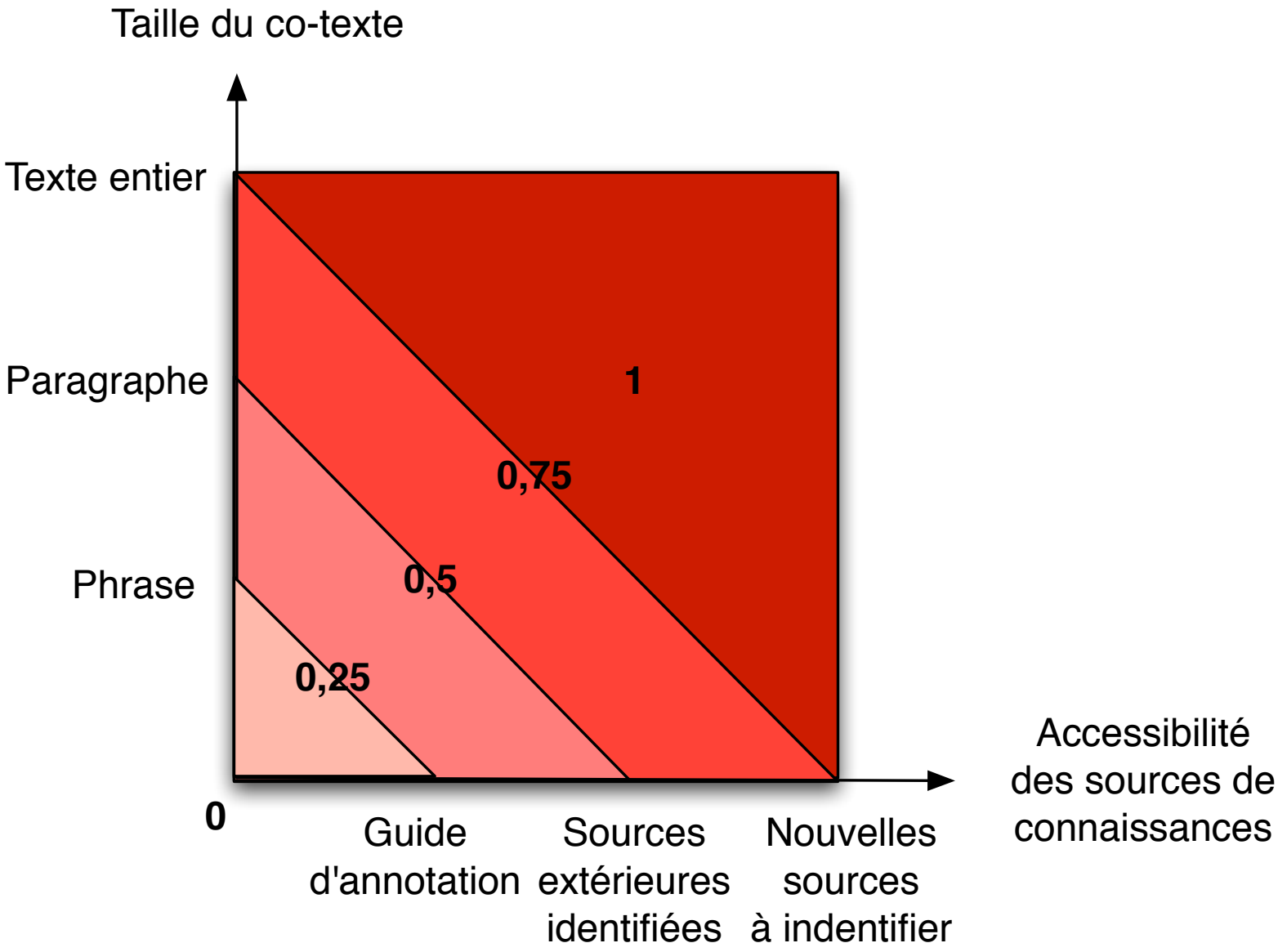
Poids du contexte

- **taille de la fenêtre** de signal source à prendre en compte :
  - La phrase :
    I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.
  - ... ou plus :



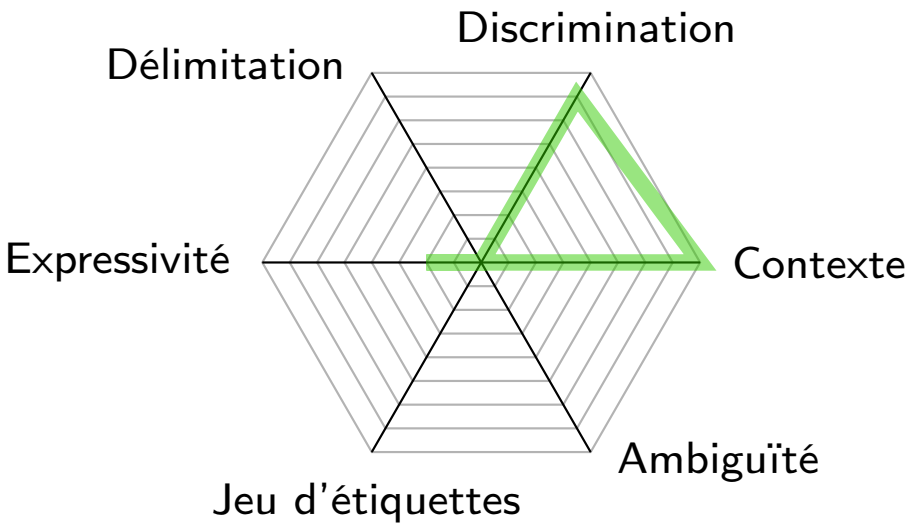
- nombre de **connaissances** à mobiliser ou degré d'accessibilité des sources de connaissances qui sont consultées :
  - guide d'annotation
  - nomenclatures (Swiss-Prot)
  - nouvelles sources à trouver (Wikipedia, etc.)

Poids du contexte

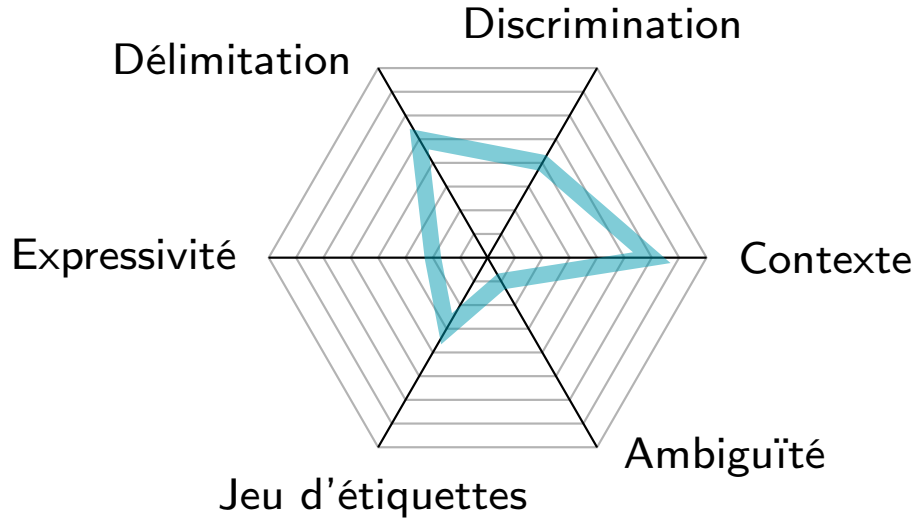


Outiller à bon escient. . .

Renommage de noms de gènes



Entités nommées structurées



. . . en fonction du profil de complexité de la campagne

Gérer la complexité de l'annotation manuelle

Quoi annoter ?

Comment annoter ?

Le poids du contexte

Outiller à bon escient

Pour finir

CQFR : Ce Qu'il Faut Retenir

Dimensions de complexité :

- discrimination
- délimitation
- expressivité du langage d'annotation
- dimension du jeu d'étiquettes
- ambiguïté
- contexte


Fort, K., François, C., Galibert, O., and Ghribi, M. (2012a).
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.


Fort, K., Nazarenko, A., and Rosset, S. (2012b).
Modeling the complexity of manual annotation tasks : a grid of analysis.
In Proceedings of the International Conference on Computational Linguistics (COLING), pages 895–910, Mumbaï, Inde.


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.

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.