



Annotation collaborative de corpus : Dimensions de complexité

Karën Fort

karen.fort@sorbonne-universite.fr / <https://www.schplaf.org/kf/>

11 décembre 2020



Gérer la complexité de l'annotation manuelle

Exemples d'annotations

Formaliser les dimensions de complexité

Quoi annoter ?

Comment annoter ?

Le poids du contexte

Outiler à bon escient

Pour finir

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 ?

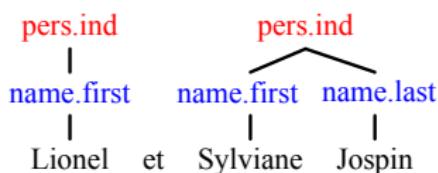
Renommage de gènes [Fort et al., 2012a] :

*The *ypdB* gene complemented the defect of the *recG40* strain. *ypdB* 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] :

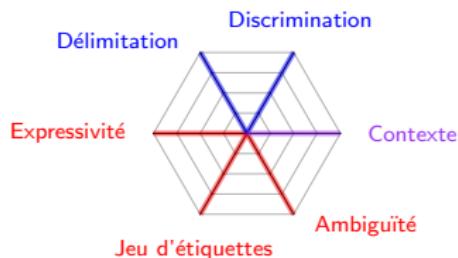


Qu'est-ce qui est complexe ?



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

Gérer la complexité de l'annotation manuelle

Quoi annoter ?

Discrimination

Délimitation des frontières

Comment annoter ?

Le poids du contexte

Outiller à bon escient

Pour finir

Gérer la complexité de l'annotation manuelle

Quoi annoter ?

Discrimination

Délimitation des frontières

Comment annoter ?

Le poids du contexte

Outiller à bon escient

Pour finir

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

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

⇒ 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.

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

Gérer la complexité de l'annotation manuelle

Quoi annoter ?

Discrimination

Délimitation des frontières

Comment annoter ?

Le poids du contexte

Outiler à bon escient

Pour finir

Délimitation des frontières

Délimiter les frontières consiste à :

- ▶ **étendre ou rétrécir** l'unité discriminée :

Madame Chirac → Monsieur et Madame Chirac

Délimitation des frontières

Délimiter les frontières consiste à :

- ▶ **étendre ou rétrécir** l'unité discriminée :
Madame Chirac → Monsieur et Madame Chirac
- ▶ **décomposer** une unité discriminée en plusieurs éléments :
le préfet Érignac → le préfet Érignac

Délimitation des frontières

Délimiter les frontières consiste à :

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

Le poids du contexte

Outiller à bon escient

Pour finir

Gérer la complexité de l'annotation manuelle

Quoi annoter ?

Comment annoter ?

Expressivité du langage d'annotation

Dimension du jeu d'étiquettes

Degré d'ambiguité

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

$$\text{Expressivité}_{Renommage} = 0,25$$

$$\text{Expressivité}_{PTB_{POS}} = 0,25$$

Gérer la complexité de l'annotation manuelle

Quoi annoter ?

Comment annoter ?

Expressivité du langage d'annotation

Dimension du jeu d'étiquettes

Degré d'ambiguité

Le poids du contexte

Outiller à bon escient

Pour finir

Dimension du jeu d'étiquettes

Person		Function				
<i>pers.ind</i> (individual person)	<i>pers.coll</i> (group of persons)	<i>func.ind</i> (individual function)	<i>func.coll</i> (collectivity of functions)			
Location			Production			
administrative (<i>loc.adm.town</i> , <i>loc.adm.reg</i> , <i>loc.adm.nat</i> , <i>loc.adm.sup</i>)	physical (<i>loc.phys.geo</i> , <i>loc.phys.hydro</i> , <i>loc.phys.astro</i>)	facilities (<i>loc.fac</i>), oronyms (<i>loc.oro</i>), address (<i>loc.add.phys</i> , <i>loc.add.elec</i>)	<i>prod.object</i> (manufactured object)	<i>prod.serv</i> (transportation route)		
			<i>prod.doctr</i> (doctrine)	<i>prod.rule</i> (law)		
			<i>prod.art</i>	<i>prod.media</i>		
Organization			Time			
<i>org.adm</i> (administration)	<i>org.ent</i> (services)		<i>time.date.abs</i> (absolute date), <i>time.date.rel</i> (relative date)	<i>time.hour.abs</i> (absolute hour), <i>time.hour.rel</i> (relative hour)		
Amount						
<i>amount</i> (with unit or general object), including duration						

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

Dimension du jeu d'étiquettes

Person			Function		
<i>pers.ind</i> (individual person)	<i>pers.coll</i> (group of persons)	<i>func.ind</i> (individual function)	<i>func.coll</i> (collectivity of functions)		
Location			Production		
administrative (<i>loc.adm.town</i> , <i>loc.adm.reg</i> , <i>loc.adm.nat</i> , <i>loc.adm.sup</i>)	physical (<i>loc.phys.geo</i> , <i>loc.phys.hydro</i> , <i>loc.phys.astro</i>)	facilities (<i>loc.fac</i>), oronyms (<i>loc.oro</i>), address (<i>loc.add.phys</i> , <i>loc.add.elec</i>)	<i>prod.object</i> (manufactured object)	<i>prod.serv</i> (transportation route)	<i>prod.fin</i> (financial products)
			<i>prod.doctr</i> (doctrine)	<i>prod.rule</i> (law)	<i>prod.soft</i> (software)
			<i>prod.art</i>	<i>prod.media</i>	<i>prod.award</i>
Organization			Time		
<i>org.adm</i> (administration)	<i>org.ent</i> (services)	<i>time.date.abs</i> (absolute date), <i>time.date.rel</i> (relative date)	<i>time.hour.abs</i> (absolute hour), <i>time.hour.rel</i> (relative hour)		
Amount					
<i>amount</i> (with unit or general object), including duration					

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

Dimension du jeu d'étiquettes

Person			Function		
<i>pers.ind</i> (individual person)	<i>pers.coll</i> (group of persons)		<i>func.ind</i> (individual function)	<i>func.coll</i> (collectivity of functions)	
Location			Production		
administrative (<i>loc.adm.town</i> , <i>loc.adm.reg</i> , <i>loc.adm.nat</i> , <i>loc.adm.sup</i>)	physical (<i>loc.phys.geo</i> , <i>loc.phys.hydro</i> , <i>loc.phys.astro</i>)	facilities (<i>loc.fac</i>), oronyms (<i>loc.oro</i>), address (<i>loc.add.phys</i> , <i>loc.add.elec</i>)	<i>prod.object</i> (manufactured object)	<i>prod.serv</i> (transportation route)	<i>prod.fin</i> (financial products)
			<i>prod.doctr</i> (doctrine)	<i>prod.rule</i> (law)	<i>prod.soft</i> (software)
			<i>prod.art</i>	<i>prod.media</i>	<i>prod.award</i>
Organization			Time		
<i>org.adm</i> (administration)	<i>org.ent</i> (services)		<i>time.date.abs</i> (absolute date), <i>time.date.rel</i> (relative date)	<i>time.hour.abs</i> (absolute hour), <i>time.hour.rel</i> (relative hour)	
Amount					
<i>amount</i> (with unit or general object), including duration					

Niveau 1 : *pers*, *func*, *loc*, *prod*, *org*, *time*, *amount* → 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* → 9 possibilités (degré de liberté = 8).

Dimension du jeu d'étiquettes

Person		Function		
<i>pers.ind</i> (individual person)	<i>pers.coll</i> (group of persons)	<i>func.ind</i> (individual function)	<i>func.coll</i> (collectivity of functions)	
Location			Production	
administrative (<i>loc.adm.town</i> , <i>loc.adm.reg</i> , <i>loc.adm.nat</i> , <i>loc.adm.sup</i>)	physical (<i>loc.phys.geo</i> , <i>loc.phys.hydro</i> , <i>loc.phys.astro</i>)	facilities (<i>loc.fac</i>), oronyms (<i>loc.oro</i>), address (<i>loc.add.phys</i> , <i>loc.add.elec</i>)	<i>prod.object</i> (manufactured object)	<i>prod.serv</i> (transportation route)
			<i>prod.doctr</i> (doctrine)	<i>prod.rule</i> (law)
			<i>prod.art</i>	<i>prod.media</i>
Organization			Time	
<i>org.adm</i> (administration)	<i>org.ent</i> (services)		<i>time.date.abs</i> (absolute date), <i>time.date.rel</i> (relative date)	<i>time.hour.abs</i> (absolute hour), <i>time.hour.rel</i> (relative hour)
Amount				
<i>amount</i> (with unit or general object), including duration				

Niveau 1 : *pers*, *func*, *loc*, *prod*, *org*, *time*, *amount* → 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* → 9 possibilités (degré de liberté = 8).

Niveau 3 : *loc.adm.town*, *loc.adm.reg*, *loc.adm.nat*, *loc.adm.sup* → 4 possibilités (degré de liberté = 3).

Dimension du jeu d'étiquettes

Degré de liberté

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

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

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

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'ambiguité

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

$$\text{Ambiguité}_{\text{Res}}(\text{Flux}) = \frac{|\text{Annotations}_{\text{amb}}|}{|\text{Annotations}|}$$

$$\text{Ambiguité}_{\text{Res}}_{\text{Renommage}} = 0,02$$

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

Degré d'ambiguité : ambiguïté théorique

Proportion des unités à annoter qui correspondent 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

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

→ Ne s'applique pas aux relations de renommage.

Gérer la complexité de l'annotation manuelle

Quoi annoter ?

Comment annoter ?

Le poids du contexte

Outiler à bon escient

Pour finir

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 :

Fabien Lévéque : C'est bien fait , avec [Gouffran] maintenant . [Gouffran] qui va tenter sa chance , et ça fait le but . Le but !

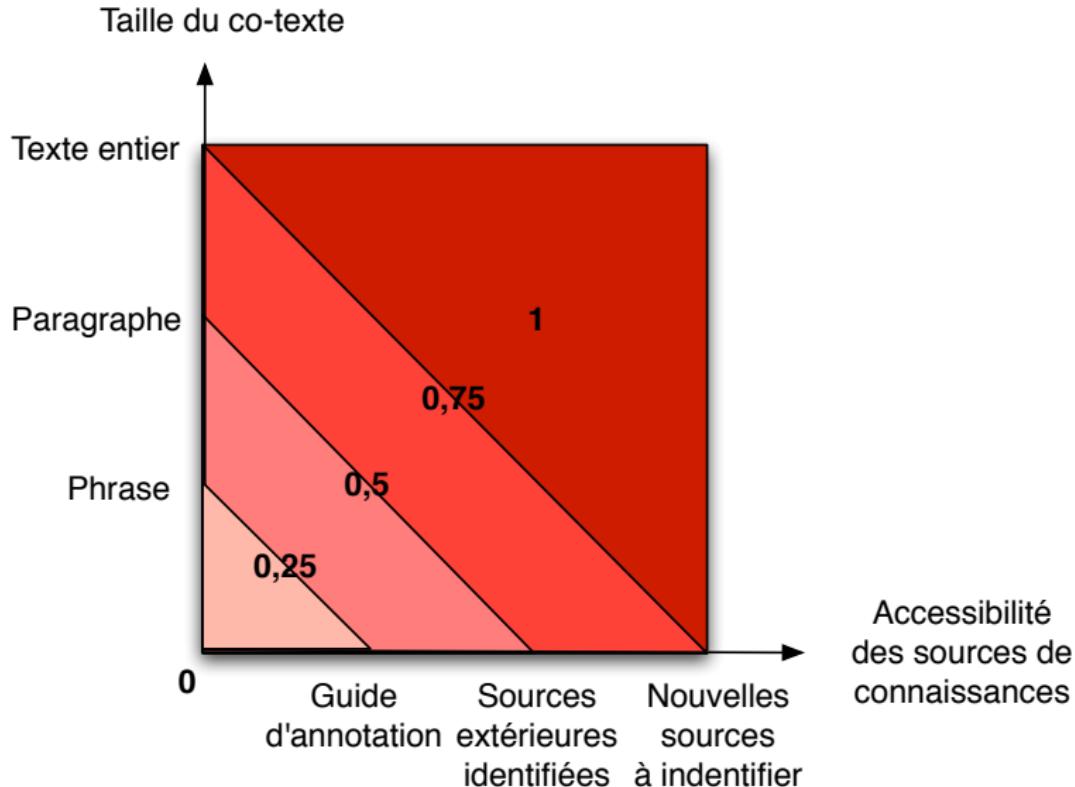
Xavier Gravelaine : Oh la la la la !

Fabien Lévéque : Et le but du plus breton des [Girondins] . C'est [Yoann Gourcuff] qui vient mettre un quatrième but ici au [stade de France] . Le cauchemar continue pour le [VOC] . Quatre à zéro en faveur des [Girondins].

(0=9 V8)

- ▶ 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



Gérer la complexité de l'annotation manuelle

Quoi annoter ?

Comment annoter ?

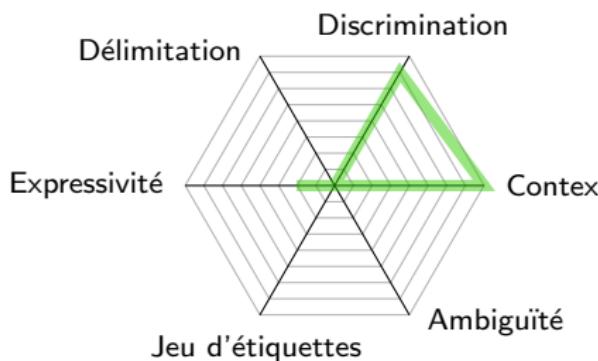
Le poids du contexte

Outiller à bon escient

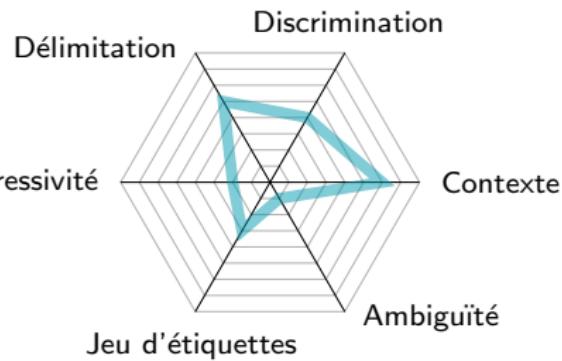
Pour finir

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, Mumbai, 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.