

Pépites (algorithmiques) pour le Traitement Automatique des Langues (TAL)

Karën Fort

karen.fort@loria.fr

4 juin 2013



1 Au cœur des systèmes, les ressources langagières

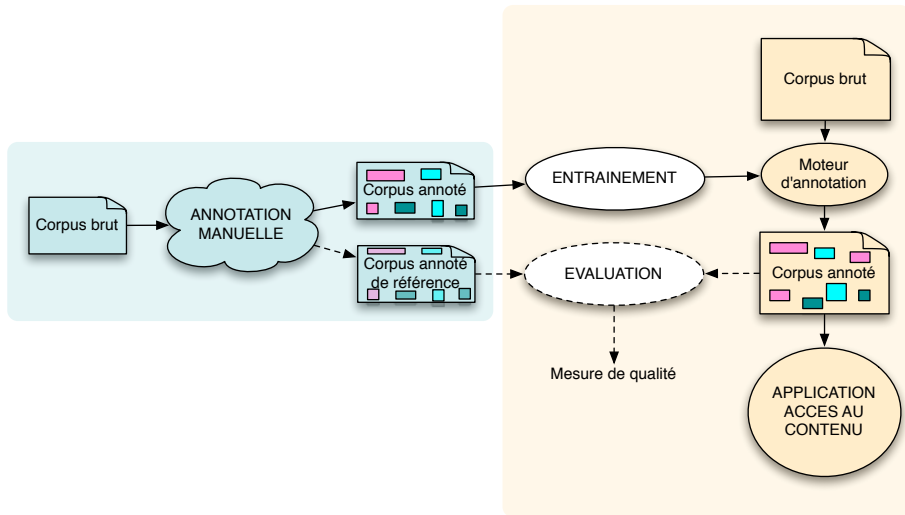
- Coût et complexité
- Solutions : expansion d'étiquettes
- Solutions : pré-annotation
- Plus loin dans la pré-annotation
- Solutions : apprentissage actif
- Solutions : formation et méthodologie

2 Focus sur le Crowdsourcing

3 Accord(s) inter-annotateurs

4 Conclusion

Annotation manuelle de corpus et TAL



Exemples d'annotation

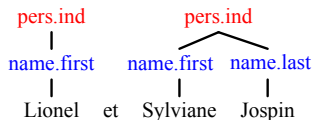
Parties du discours (*Penn Treebank*, [Marcus et al., 1993]), pré-annotées :

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

Renommage de gènes [LREC 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_(recU) and ypbC_(recS) 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.

Entités nommées structurées [LAW 2011] :



Un coût élevé, voire prohibitif

Penn Treebank [Marcus et al., 1993] :

- 4,8 millions de tokens annotés en morpho-syntaxe $\Rightarrow$ phase d'apprentissage d'1 mois, pour atteindre 3 000 mots/h
- 3 millions de tokens annotés en syntaxe $\Rightarrow$ phase d'apprentissage d'2 mois, pour atteindre 475 mots/h

Prague Dependency Treebank [Böhmová et al., 2001] :

- 1,8 millions de tokens annotés en morpho-syntaxe et syntaxe
- $\Rightarrow$ 5 ans, 22 personnes (max. 17 en parallèle), 600 000 dollars

GENIA [Kim et al., 2008] :

- 9 372 phrases annotés en microbiologie (noms de gènes, protéines)
- $\Rightarrow$ 5 annotateurs à temps partiel, 1 coordinateur senior et 1 junior pendant 1,5 an

Solutions ?

Active Learning

Tag Dictionaries

Pre-annotation

Annotation Tools

Crowdsourcing

Methodology

Training

1 Au cœur des systèmes, les ressources langagières

- Coût et complexité
- **Solutions : expansion d'étiquettes**
- Solutions : pré-annotation
- Plus loin dans la pré-annotation
- Solutions : apprentissage actif
- Solutions : formation et méthodologie

2 Focus sur le Crowdsourcing

3 Accord(s) inter-annotateurs

4 Conclusion

Expansion d'étiquettes (*Tag Dictionaries*)

Permet de :

- ➊ stocker les catégories associées par un annotateurs à un token
- ➋ proposer ces catégories lorsqu'on rencontre le même token

⇒ Très **simple** et relativement efficace (voir [Carmen et al., 2010]), mais la méthode est d'autant plus efficace qu'on a déjà beaucoup annoté.

1 Au cœur des systèmes, les ressources langagières

- Coût et complexité
- Solutions : expansion d'étiquettes
- **Solutions : pré-annotation**
- Plus loin dans la pré-annotation
- Solutions : apprentissage actif
- Solutions : formation et méthodologie

2 Focus sur le Crowdsourcing

3 Accord(s) inter-annotateurs

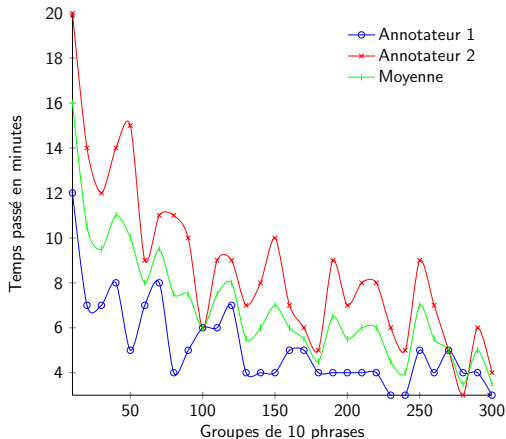
4 Conclusion

Annotation morpho-syntaxique du *Penn Treebank*

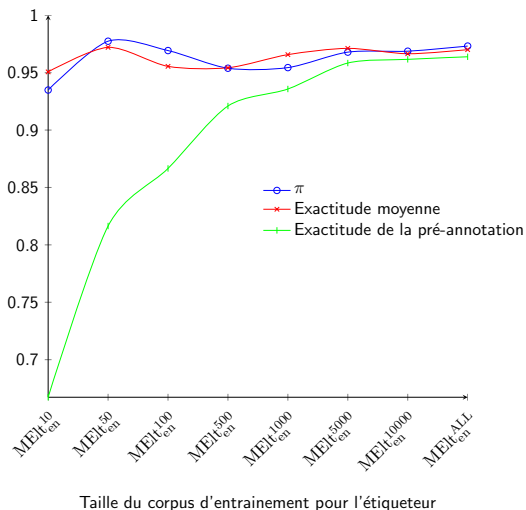
I/PRP do/VBP n't/RB feel/VB very/RB ferocious/JJ ./.

→ tester l'influence de la pré-annotation [Fort and Sagot, 2010]

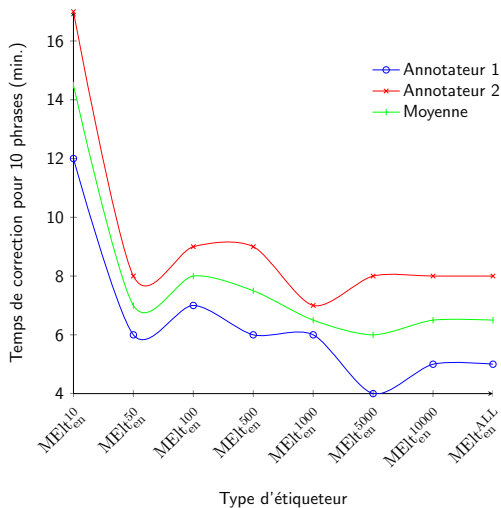
Courbe d'apprentissage : annotation morpho-syntaxique du *Penn Treebank*



Exactitude de l'annotation et π en fonction du type de pré-annotation



Temps d'annotation en fonction de la qualité de la pré-annotation



Exercice : transcrire de la parole

Transcriber

Notez vos questions, vos décisions.

Sources

Le fichier audio que vous allez transcrire est la propriété du projet Quæro.

- Il n'est **pas** librement disponible et utilisable

⇒ Vous devrez le supprimer à la fin du cours

Transcription manuelle vs assistée [Bazillon et al., 2008]

	Parole préparée	Parole spontanée
Transcription manuelle	17 h 36 min	19 h 33 min
Transcription assistée	8 h 31 min	15 h 44 min

TABLE: Temps de transcription pour 2h de parole

	Parole préparée	Parole spontanée
Transcription manuelle	16,95	35,21
Transcription assistée	15,83	34,33

TABLE: Word Error Rate

Plus de détails

	Parole préparée	Parole spontanée
Transcription manuelle	13 h 36 min	16 h 15 min
Transcription assistée	5 h 06 min	12 h 41 min

TABLE: Temps de transcription et de segmentation

	Parole préparée	Parole spontanée
Transcription manuelle	2 h 43 min	1 h 05 min
Transcription assistée	2h 08 min	0 h 51 min

TABLE: Temps de correction

1 Au cœur des systèmes, les ressources langagières

- Coût et complexité
- Solutions : expansion d'étiquettes
- Solutions : pré-annotation
- Plus loin dans la pré-annotation
- **Solutions : apprentissage actif**
- Solutions : formation et méthodologie

2 Focus sur le Crowdsourcing

3 Accord(s) inter-annotateurs

4 Conclusion

Cas particulier : apprentissage actif (active learning)

- Toutes les annotations ne sont pas nécessaires pour entraîner un outil
⇒ détecter les annotations qui sont réellement utiles pour améliorer les résultats finaux
- Pré-annoter un corpus automatiquement, puis demander aux annotateurs de corriger, ré-annoter, etc.

⇒ itératif

+ permet de **gagner du temps**

- mais **consommateur de temps** si le système est trop mauvais (à définir)

- sur le projet Ritel (dialogue oral humain-machine) : plus de 30 % d'erreurs, il était **plus rapide** pour les transpositeurs de transcrire directement

1 Au cœur des systèmes, les ressources langagières

- Coût et complexité
- Solutions : expansion d'étiquettes
- Solutions : pré-annotation
- Plus loin dans la pré-annotation
- Solutions : apprentissage actif
- Solutions : formation et méthodologie

2 Focus sur le Crowdsourcing

3 Accord(s) inter-annotateurs

4 Conclusion

Formation et documentation

Une bonne formation des annotateurs est la meilleure solution pour une annotation de bonne qualité et rapide [Dandapat et al., 2009].

Cette formation doit être associée à une documentation adaptée avec :

- une définition claire de l'**application**
- une définition claire et détaillée des **catégories** (toujours possible ou même désirable ?)
- des **exemples** significatifs
- **catégories ambiguës** présentées en parallèle, comme dans la documentation du PTB (voir : [ftp ://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz](ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz))

Formation et documentation

Une bonne formation des annotateurs est la meilleure solution pour une annotation de bonne qualité et rapide [Dandapat et al., 2009].

Cette formation doit être associée à une documentation adaptée avec :

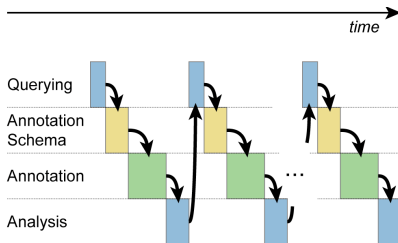
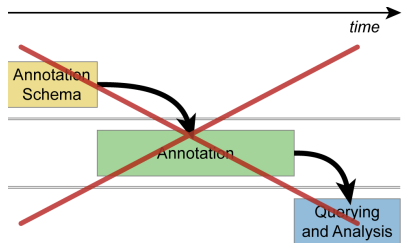
- une définition claire de l'**application**
- une définition claire et détaillée des **catégories** (toujours possible ou même désirable ?)
- des **exemples** significatifs
- **catégories ambiguës** présentées en parallèle, comme dans la documentation du PTB (voir :
[ftp ://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz](ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz))

Il faut garder à l'esprit que les annotateurs sont **au cœur** de la campagne d'annotation !

Méthodologie

- 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** pendant l'annotation, pour vérifier que les annotateurs sont cohérents avec eux-mêmes [Gut and Bayerl, 2004].
- Cela peut aller jusqu'à l'**annotation agile** [Voormann and Gut, 2008, Alex et al., 2010], qui implique plusieurs itérations

Annotation agile



- 1 Au cœur des systèmes, les ressources langagières
- 2 **Focus sur le Crowdsourcing**
 - Définition
 - Jeux sérieux
 - Myriadisation du travail parcellisé
 - La légende de la foule et des non-experts
- 3 Accord(s) inter-annotateurs
- 4 Conclusion

Crowdsourcing : définition

Crowdsourcing est le fait de délocaliser des tâches, traditionnellement réalisées par des employés ou des contractants, vers un large groupe de personnes indéterminé (une foule), à travers un appel ouvert.

The term "crowdsourcing" is a portmanteau of "crowd" and "outsourcing", first coined by Jeff Howe in a June 2006 Wired magazine article "The Rise of Crowdsourcing". Howe explains that because technological advances have allowed for cheap consumer electronics, the gap between professionals and amateurs has been diminished. Companies are then able to take advantage of the talent of the public, and Howe states that "It's not outsourcing; it's crowdsourcing."

(Wikipedia, consultée le 2 déc. 2010)

Différent types de crowdsourcing

Lié au Web 2.0 :

- **crowdvoting** : utiliser les réseaux sociaux pour voter sur une question, un produit, etc (*social bookmarking*)
- **crowdcreation** : compétition d'idées
- **crowdwisdom** : réponses à des questions (Yahoo! questions)
- **crowdfunding** : pour des projets artistiques, des campagnes politiques, etc.

... à travers les réseaux sociaux, les jeux sérieux et la myriadisation du travail parcellisé.

Différent types de crowdsourcing

Lié au Web 2.0 :

- **crowdvoting** : utiliser les réseaux sociaux pour voter sur une question, un produit, etc (*social bookmarking*)
- **crowdcreation** : compétition d'idées
- **crowdwisdom** : réponses à des questions (Yahoo! questions)
- **crowdfunding** : pour des projets artistiques, des campagnes politiques, etc.

... à travers les ~~réseaux sociaux~~, les **jeux sérieux** et la **myriadisation du travail parcellisé**.

- 1 Au cœur des systèmes, les ressources langagières
- 2 **Focus sur le Crowdsourcing**
 - Définition
 - **Jeux sérieux**
 - Myriadisation du travail parcellisé
 - La légende de la foule et des non-experts
- 3 Accord(s) inter-annotateurs
- 4 Conclusion

Crowdsourcing : jeux sérieux

- **ESP game** : 13 500 utilisateurs ont étiqueté 1,3 millions d'images en 3 mois! [von Ahn, 2006]
- [JeuxDeMots](#) [Lafourcade, 2007]
- [PhraseDetectives](#) [Chamberlain et al., 2008]

Crowdsourcing : le monde selon GWAP (jeu ayant un but)



USERPROFILE

kaa

- 0 this week
- 0 decisions
- 0 agreements
- 0 extras

0 this month

31 all time

Level: Trainee

Your rating: **80%**

CASE OPEN
50 tasks remaining

1 completed case

[EDIT PROFILE](#) | [LOGOUT](#)

 Instructions

 FAQ

 Restart

DETECTIVES CONFERENCE

Another detective has made a decision about a phrase, either that it refers to another phrase, it has not been mentioned before, it is a property or it does not refer to anything. Do you agree with them?




The Adventures of Sherlock Holmes - The Man with the Twisted Lip (Sir Arthur Conan Doyle)

"Certainly."

"Then dress. No one is stirring yet, but I know where the stable-boy sleeps, and we shall soon have the trap out." He chuckled to himself as he spoke, his eyes twinkled, and he seemed a different man to the sombre thinker of the previous night.

As I dressed I glanced at my watch. It was no wonder that no one was stirring. It was twenty-five minutes past four. I had hardly finished when Holmes returned with the news that the boy was putting in the horse.

"I want to test a little theory of mine," said he, pulling on his boots. "I think, Watson, that you are now standing in the presence of one of the most absolute fools in Europe. I deserve to be kicked from here to Charing Cross. But I think I have the key of the affair now."

"And where is it?" I asked, smiling.

"In the bathroom," he answered. "Oh, yes, I am not joking," he continued, seeing my look of incredulity. "I have just been there, and I have taken it out, and I have got it in this Gladstone bag. Come on, **my boy**, and **we** shall see whether it will not fit the lock."

The phrase in blue is the closest phrase that refers to the phrase in orange.

 Disagree

 Agree

PHRASE DETECTIVES

SEARCHCLUES

Words like **they**, him, her and it are likely to refer to something else in the text. Try to find the closest mention of this phrase.

Words like **they** or **them** could refer to more than one thing in the text so select more than one phrase if necessary.

Always look for the closest previous mention of the phrase to score maximum agreement points.

Feedback 

Crowdsourcing : questions

[Fort et al., 2011, Chamberlain et al., 2012] :

- qualité ?
- triche ?
- droits d'auteur ?
- éthique ?

- 1 Au cœur des systèmes, les ressources langagières
- 2 **Focus sur le Crowdsourcing**
 - Définition
 - Jeux sérieux
 - **Myriadisation du travail parcellisé**
 - La légende de la foule et des non-experts
- 3 Accord(s) inter-annotateurs
- 4 Conclusion

MTurk : Gold Mine or Coal Mine ? [Fort et al., 2011]

Présentation, LTC 2011 [Adda et al., 2011]

1 Au cœur des systèmes, les ressources langagières

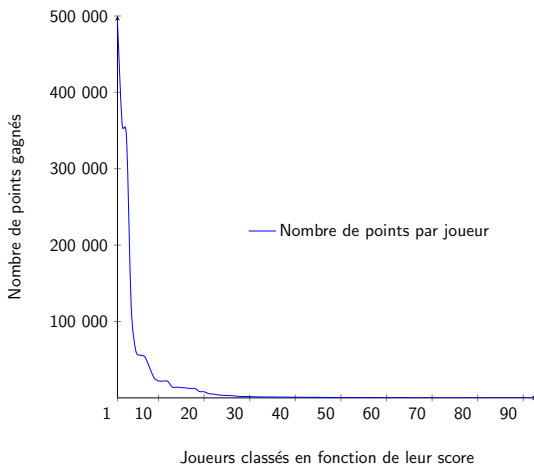
2 Focus sur le Crowdsourcing

- Définition
- Jeux sérieux
- Myriadisation du travail parcellisé
- La légende de la foule et des non-experts

3 Accord(s) inter-annotateurs

4 Conclusion

Crowdsourcing : une foule de "non-experts" ? (GWAP)



Nombre de joueurs sur *Phrase Detectives* [Chamberlain et al., 2012] en fonction de leur classement en points (février 2011 - février 2012)

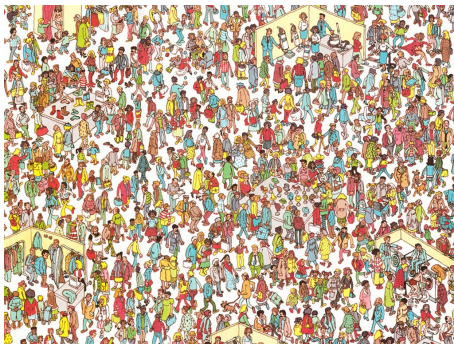
Crowdsourcing : une foule de "non-experts" ? (MTurk)

Nombre de *Turkers* actifs sur MTurk :

- nombre enregistré sur le site : plus de 500 000
- réellement actifs : entre 15 059 et 42 912 [Fort et al., 2011]
- 80 % des tâches (HIT) sont réalisées par les 20 % de *Turkers* les plus actifs [Deneme, 2009]

Crowdsourcing

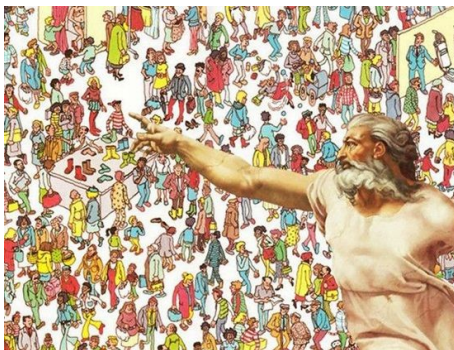
Faire annoter des "non-experts" ?



Crowdsourcing

Faire annoter des "non-experts" ?

→ Trouver des experts dans la foule



Crowdsourcing : avantages et inconvénients

- + accès à plus de monde
 - qui travaille ? (langue maternelle ? Formation ?)
- + coût faible (quand ce n'est pas gratuit)
 - des jeux au hobby à ... l'exploitation !
- + rapide
- + bonne qualité [Snow et al., 2008]...
 - ... si l'annotation est peu complexe !
 - et si les gens ne trichent pas !

- 1 Au cœur des systèmes, les ressources langagières
- 2 Focus sur le Crowdsourcing
- 3 **Accord(s) inter-annotateurs**
 - Exercice
 - Prendre en compte le hasard
 - Coefficient S
 - Coefficient π
 - Coefficient κ
 - Kappa et annotables
 - Interprétation
- 4 Conclusion

Validité vs fiabilité [Artstein and Poesio, 2008]

- On s'intéresse à la **validité** de l'annotation manuelle

Validité vs fiabilité [Artstein and Poesio, 2008]

- On s'intéresse à la **validité** de l'annotation manuelle
 - ▶ c-à-d si les catégories annotées sont correctes

Validité vs fiabilité [Artstein and Poesio, 2008]

- On s'intéresse à la **validité** de l'annotation manuelle
 - ▶ c-à-d si les catégories annotées sont correctes
- Mais il n'existe pas de "vérité terrain"

Validité vs fiabilité [Artstein and Poesio, 2008]

- On s'intéresse à la **validité** de l'annotation manuelle
 - ▶ c-à-d si les catégories annotées sont correctes
- Mais il n'existe pas de "vérité terrain"
 - ▶ Les catégories sont déterminées par un jugement humain et l'annotation est une interprétation

Validité vs fiabilité [Artstein and Poesio, 2008]

- On s'intéresse à la **validité** de l'annotation manuelle
 - ▶ c-à-d si les catégories annotées sont correctes
- Mais il n'existe pas de "vérité terrain"
 - ▶ Les catégories sont déterminées par un jugement humain et l'annotation est une interprétation
 - ▶ Conséquence : impossible de mesurer directement la validité d'une annotation

Validité vs fiabilité [Artstein and Poesio, 2008]

- On s'intéresse à la **validité** de l'annotation manuelle
 - ▶ c-à-d si les catégories annotées sont correctes
- Mais il n'existe pas de "vérité terrain"
 - ▶ Les catégories sont déterminées par un jugement humain et l'annotation est une interprétation
 - ▶ Conséquence : impossible de mesurer directement la validité d'une annotation
- À la place, on mesure la **fiabilité** de l'annotation

Validité vs fiabilité [Artstein and Poesio, 2008]

- On s'intéresse à la **validité** de l'annotation manuelle
 - ▶ c-à-d si les catégories annotées sont correctes
- Mais il n'existe pas de "vérité terrain"
 - ▶ Les catégories sont déterminées par un jugement humain et l'annotation est une interprétation
 - ▶ Conséquence : impossible de mesurer directement la validité d'une annotation
- À la place, on mesure la **fiabilité** de l'annotation
 - ▶ c-à-d si les annotateurs humains prennent des décisions **cohérentes** ⇒ ils ont assimilé le guide d'annotation

Validité vs fiabilité [Artstein and Poesio, 2008]

- On s'intéresse à la **validité** de l'annotation manuelle
 - ▶ c-à-d si les catégories annotées sont correctes
- Mais il n'existe pas de "vérité terrain"
 - ▶ Les catégories sont déterminées par un jugement humain et l'annotation est une interprétation
 - ▶ Conséquence : impossible de mesurer directement la validité d'une annotation
- À la place, on mesure la **fiabilité** de l'annotation
 - ▶ c-à-d si les annotateurs humains prennent des décisions **cohérentes** ⇒ ils ont assimilé le guide d'annotation
 - ▶ Hypothèse : cohérence élevée implique validité

Validité vs fiabilité [Artstein and Poesio, 2008]

- On s'intéresse à la **validité** de l'annotation manuelle
 - ▶ c-à-d si les catégories annotées sont correctes
- Mais il n'existe pas de "vérité terrain"
 - ▶ Les catégories sont déterminées par un jugement humain et l'annotation est une interprétation
 - ▶ Conséquence : impossible de mesurer directement la validité d'une annotation
- À la place, on mesure la **fiabilité** de l'annotation
 - ▶ c-à-d si les annotateurs humains prennent des décisions **cohérentes** ⇒ ils ont assimilé le guide d'annotation
 - ▶ Hypothèse : cohérence élevée implique validité
- Comment déterminer la fiabilité ?

Fiabilité (cohérence)

- chaque unité est annotée par un seul annotateur, et on réalise des vérifications aléatoires ($\approx$ seconde annotation)
- certaines unités sont annotées par plusieurs annotateurs
- chaque unité est annotée par plusieurs annotateurs - suivi d'une discussion sur les désaccords
- chaque unité est annotée par plusieurs annotateurs - suivi d'une adjudication (expert)

Dans tous les cas, mesure de fiabilité : [coefficients d'accord inter-annotateurs](#)

- 1 Au cœur des systèmes, les ressources langagières
- 2 Focus sur le Crowdsourcing
- 3 **Accord(s) inter-annotateurs**
 - Exercice
 - Prendre en compte le hasard
 - Coefficient S
 - Coefficient π
 - Coefficient κ
 - Kappa et annotables
 - Interprétation
- 4 Conclusion

Exemple

Validation d'annotations sémantiques (contenu/contenant) :

Phrase	A	B	D'accord ?
Put tea in a heat-resistant jug and add the boiling water.	✓	✓	✓
Where are the batteries kept in a phone ?	✗	✓	✗
Vinegar's usefulness doesn't stop inside the house .	✗	✗	✓
How do I recognize a room that contains radioactive materials ?	✓	✓	✓
A letterbox is a plastic, screw-top bottle that contains a small notebook and a unique rubber stamp.	✓	✗	✗

→ **Accord inter-annotateurs ?**

Représentation synthétique

		A		
		✓	✗	Total
B	✓	4	2	6
	✗	2	2	4
	Total	6	4	10

Accord observé (A_o)

proportion de réponses sur lesquelles les annotateurs sont du même avis.

Ici :

Représentation synthétique

		A		
		✓	✗	Total
B	✓	4	2	6
	✗	2	2	4
	Total	6	4	10

Accord observé (A_o)

proportion de réponses sur lesquelles les annotateurs sont du même avis.

Ici : $A_o = \frac{4+2}{10} = \mathbf{0,6}$

- 1 Au cœur des systèmes, les ressources langagières
- 2 Focus sur le Crowdsourcing
- 3 **Accord(s) inter-annotateurs**
 - Exercice
 - **Prendre en compte le hasard**
 - Coefficient S
 - Coefficient π
 - Coefficient κ
 - Kappa et annotables
 - Interprétation
- 4 Conclusion

Et si...

... une partie de l'accord était due au **hasard** :

dans notre exemple, quelle proportion d'accord peut être due au hasard ?

Et si...

... une partie de l'accord était due au **hasard** :

- Deux annotateurs annotant au hasard seront d'accord **la moitié du temps** (0,5).
- L'accord qui peut être dû au hasard varie selon le schéma d'annotation et les données annotées.

L'accord significatif est celui qui se situe **au-dessus du hasard**.

→ rejoint le concept de *baseline* (résultat plancher).

Prendre en compte le hasard

Accord attendu (Expected Agreement : A_e)

valeur attendue de l'accord observé.

Accord au-dessus du hasard : $A_o - A_e$

Accord maximum possible au-dessus du hasard : $1 - A_e$

Proportion d'accord au-dessus du hasard atteinte : $\frac{A_o - A_e}{1 - A_e}$

Accord parfait : $\frac{1 - A_e}{1 - A_e}$

Désaccord parfait : $\frac{-A_e}{1 - A_e}$

Accord attendu (expected agreement)

Comment calculer l'accord attendu par hasard
(A_e) ?

- 1 Au cœur des systèmes, les ressources langagières
- 2 Focus sur le Crowdsourcing
- 3 **Accord(s) inter-annotateurs**
 - Exercice
 - Prendre en compte le hasard
 - **Coefficient S**
 - Coefficient π
 - Coefficient κ
 - Kappa et annotables
 - Interprétation
- 4 Conclusion

S [Bennett et al., 1954]

S

Même probabilité pour tous les annotateurs et toutes les catégories.

Nombre de catégories : q

Probabilité qu'un annotateur choisisse une catégorie $q_a : \frac{1}{q}$

Probabilité que les deux annotateurs choisissent une catégorie $q_a : (\frac{1}{q})^2$

Probabilité que les deux annotateurs choisissent la même catégorie :

$$A_e^S = q \cdot (\frac{1}{q})^2 = \frac{1}{q}$$

Toutes les catégories sont considérées de la même manière : conséquences

	Yes	No	Total
Yes	20	5	25
No	5	20	25
Total	25	25	50

Toutes les catégories sont considérées de la même manière : conséquences

	Yes	No	Total
Yes	20	5	25
No	5	20	25
Total	25	25	50

$$A_o = \frac{20+20}{50} = 0,8$$

$$A_e^S = \frac{1}{2} = 0,5$$

$$S = \frac{0,8-0,5}{1-0,5} = \mathbf{0,6}$$

Toutes les catégories sont considérées de la même manière : conséquences

	Yes	No	Total
Yes	20	5	25
No	5	20	25
Total	25	25	50

	Yes	No	C	D	Total
Yes	20	5	0	0	25
No	5	20	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	25	25	0	0	50

$$A_o = \frac{20+20}{50} = 0,8$$

$$A_e^S = \frac{1}{2} = 0,5$$

$$S = \frac{0,8-0,5}{1-0,5} = \mathbf{0,6}$$

Toutes les catégories sont considérées de la même manière : conséquences

	Yes	No	Total
Yes	20	5	25
No	5	20	25
Total	25	25	50

	Yes	No	C	D	Total
Yes	20	5	0	0	25
No	5	20	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	25	25	0	0	50

$$A_o = \frac{20+20}{50} = 0,8$$

$$A_e^S = \frac{1}{2} = 0,5$$

$$S = \frac{0,8-0,5}{1-0,5} = \mathbf{0,6}$$

$$A_o = \frac{20+20}{50} = 0,8$$

$$A_e^S = \frac{1}{4} = 0,25$$

$$S = \frac{0,8-0,25}{1-0,25} = \mathbf{0,73}$$

- 1 Au cœur des systèmes, les ressources langagières
- 2 Focus sur le Crowdsourcing
- 3 **Accord(s) inter-annotateurs**
 - Exercice
 - Prendre en compte le hasard
 - Coefficient S
 - **Coefficient π**
 - Coefficient κ
 - Kappa et annotables
 - Interprétation
- 4 Conclusion

π [Scott, 1955] π

Des probabilités qui diffèrent selon les catégories.

Nombre total de jugements : N

Probabilité qu'un annotateur choisisse une catégorie q_a : $\frac{n_{q_a}}{N}$

Probabilité que les deux annotateurs choisissent une catégorie q_a : $(\frac{n_{q_a}}{N})^2$

Probabilité que les deux annotateurs choisissent la même catégorie :

$$A_e^\pi = \sum_q \left(\frac{n_q}{N}\right)^2 = \frac{1}{N^2} \sum_q n_q^2$$

Comparer S et π

	Yes	No	Total
Yes	20	5	25
No	5	20	25
Total	25	25	50

	Yes	No	C	D	Total
Yes	20	5	0	0	25
No	5	20	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	25	25	0	0	50

$$A_o = 0,8$$

$$S = \mathbf{0,6}$$

$$A_o = 0,8$$

$$S = \mathbf{0,73}$$

Comparer S et π

	Yes	No	Total
Yes	20	5	25
No	5	20	25
Total	25	25	50

	Yes	No	C	D	Total
Yes	20	5	0	0	25
No	5	20	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	25	25	0	0	50

$$A_o = 0,8$$

$$S = \mathbf{0,6}$$

$$A_e^\pi = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0,5$$

$$\pi = \frac{0,8-0,5}{1-0,5} = \mathbf{0,6}$$

$$A_o = 0,8$$

$$S = \mathbf{0,73}$$

Comparer S et π

	Yes	No	Total
Yes	20	5	25
No	5	20	25
Total	25	25	50

	Yes	No	C	D	Total
Yes	20	5	0	0	25
No	5	20	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	25	25	0	0	50

$$A_o = 0,8$$

$$S = \mathbf{0,6}$$

$$A_e^\pi = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0,5$$

$$\pi = \frac{0,8-0,5}{1-0,5} = \mathbf{0,6}$$

$$A_o = 0,8$$

$$S = \mathbf{0,73}$$

$$A_e^\pi = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0,5$$

$$\pi = \frac{0,8-0,5}{1-0,5} = \mathbf{0,6}$$

- 1 Au cœur des systèmes, les ressources langagières
- 2 Focus sur le Crowdsourcing
- 3 Accord(s) inter-annotateurs**
 - Exercice
 - Prendre en compte le hasard
 - Coefficient S
 - Coefficient π
 - Coefficient κ**
 - Kappa et annotables
 - Interprétation
- 4 Conclusion

κ [Cohen, 1960] κ

Différent annotateurs ont différentes interprétations des instructions (biais/préjugés). κ prend en compte les biais individuels.

Nombre total d'unités : i

Probabilité qu'un annotateur A_x choisisse une catégorie q_a : $\frac{n_{A_x q_a}}{i}$

Probabilité que les deux annotateurs choisissent une catégorie q_a :
 $\frac{n_{A_1 q_a}}{i} \cdot \frac{n_{A_2 q_a}}{i}$

Probabilité que les deux annotateurs choisissent la même catégorie :

$$A_e^\kappa = \sum_q \frac{n_{A_1 q}}{i} \cdot \frac{n_{A_2 q}}{i} = \frac{1}{i^2} \sum_q n_{A_1 q} n_{A_2 q}$$

Comparer π et κ

	Yes	No	Total
Yes	20	5	25
No	5	20	25
Total	25	25	50

	Yes	No	C	D	Total
Yes	20	5	0	0	25
No	5	20	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	25	25	0	0	50

$$A_o = 0,8$$

$$A_e^\pi = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0,5$$

$$\pi = \frac{0,8-0,5}{1-0,5} = \mathbf{0,6}$$

$$A_o = 0,8$$

$$A_e^\pi = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0,5$$

$$\pi = \frac{0,8-0,5}{1-0,5} = \mathbf{0,6}$$

Comparer π et κ

	Yes	No	Total
Yes	20	5	25
No	5	20	25
Total	25	25	50

	Yes	No	C	D	Total
Yes	20	5	0	0	25
No	5	20	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	25	25	0	0	50

$$A_o = 0,8$$

$$A_e^\pi = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0,5$$

$$\pi = \frac{0,8-0,5}{1-0,5} = \mathbf{0,6}$$

$$A_e^\kappa = \frac{(\frac{25 \times 25}{50}) + (\frac{25 \times 25}{50})}{50} = 0,5$$

$$\kappa = \frac{0,8-0,5}{1-0,5} = \mathbf{0,6}$$

$$A_o = 0,8$$

$$A_e^\pi = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0,5$$

$$\pi = \frac{0,8-0,5}{1-0,5} = \mathbf{0,6}$$

Comparer π et κ

	Yes	No	Total
Yes	20	5	25
No	5	20	25
Total	25	25	50

	Yes	No	C	D	Total
Yes	20	5	0	0	25
No	5	20	0	0	25
C	0	0	0	0	0
D	0	0	0	0	0
Total	25	25	0	0	50

$$A_o = 0,8$$

$$A_e^\pi = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0,5$$

$$\pi = \frac{0,8-0,5}{1-0,5} = \mathbf{0,6}$$

$$A_e^\kappa = \frac{(\frac{25 \times 25}{50}) + (\frac{25 \times 25}{50})}{50} = 0,5$$

$$\kappa = \frac{0,8-0,5}{1-0,5} = \mathbf{0,6}$$

$$A_o = 0,8$$

$$A_e^\pi = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0,5$$

$$\pi = \frac{0,8-0,5}{1-0,5} = \mathbf{0,6}$$

$$A_e^\kappa = \frac{(\frac{25 \times 25}{50}) + (\frac{25 \times 25}{50})}{50} = 0,5$$

$$\kappa = \frac{0,8-0,5}{1-0,5} = \mathbf{0,6}$$

Comparer π et κ

	Yes	No	Total
Yes	20	5	25
No	5	20	25
Total	25	25	50

	Yes	No	Total
Yes	24	8	32
No	14	24	38
Total	38	32	70

$$A_o = 0,8$$

$$A_e^\pi = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0,5$$

$$\pi = \frac{0,8-0,5}{1-0,5} = \mathbf{0,6}$$

$$A_o = 0,68$$

$$A_e^\pi = \frac{((\frac{38+32}{2})^2 + (\frac{32+38}{2})^2)}{70^2} = 0,5$$

$$\pi = \frac{0,68-0,5}{1-0,5} = \mathbf{0,36}$$

Comparer π et κ

	Yes	No	Total
Yes	20	5	25
No	5	20	25
Total	25	25	50

	Yes	No	Total
Yes	24	8	32
No	14	24	38
Total	38	32	70

$$A_o = 0,8$$

$$A_e^\pi = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0,5$$

$$\pi = \frac{0,8-0,5}{1-0,5} = \mathbf{0,6}$$

$$A_e^\kappa = \frac{(\frac{25 \times 25}{50}) + (\frac{25 \times 25}{50})}{50} = 0,5$$

$$\kappa = \frac{0,8-0,5}{1-0,5} = \mathbf{0,6}$$

$$A_o = 0,68$$

$$A_e^\pi = \frac{((\frac{38+32}{2})^2 + (\frac{32+38}{2})^2)}{70^2} = 0,5$$

$$\pi = \frac{0,68-0,5}{1-0,5} = \mathbf{0,36}$$

Comparer π et κ

	Yes	No	Total
Yes	20	5	25
No	5	20	25
Total	25	25	50

	Yes	No	Total
Yes	24	8	32
No	14	24	38
Total	38	32	70

$$A_o = 0,8$$

$$A_e^\pi = \frac{((\frac{25+25}{2})^2 + (\frac{25+25}{2})^2)}{50^2} = 0,5$$

$$\pi = \frac{0,8-0,5}{1-0,5} = \mathbf{0,6}$$

$$A_e^\kappa = \frac{(\frac{25 \times 25}{50}) + (\frac{25 \times 25}{50})}{50} = 0,5$$

$$\kappa = \frac{0,8-0,5}{1-0,5} = \mathbf{0,6}$$

$$A_o = 0,68$$

$$A_e^\pi = \frac{((\frac{38+32}{2})^2 + (\frac{32+38}{2})^2)}{70^2} = 0,5$$

$$\pi = \frac{0,68-0,5}{1-0,5} = \mathbf{0,36}$$

$$A_e^\kappa = \frac{(\frac{38 \times 32}{70}) + (\frac{32 \times 38}{70})}{70} = 0,49$$

$$\kappa = \frac{0,68-0,49}{1-0,49} = \mathbf{0,37}$$

S , π et κ

Dans tous les cas :

$$\begin{array}{ll} A_e^\pi \geq A_e^S & \pi \leq S \\ A_e^\pi \geq A_e^\kappa & \pi \leq \kappa \end{array}$$

- Comment définir les “annotables” ?
- Qu'est-ce qu'un “bon” κ (ou π ou S) ?

- 1 Au cœur des systèmes, les ressources langagières
- 2 Focus sur le Crowdsourcing
- 3 **Accord(s) inter-annotateurs**
 - Exercice
 - Prendre en compte le hasard
 - Coefficient S
 - Coefficient π
 - Coefficient κ
 - **Kappa et annotables**
 - Interprétation
- 4 Conclusion

Renommage de gènes

"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. <former-name1 id="1">yppB</former-name1> and <former-name2 id="2">ypbC</former-name2> and their respective null alleles were termed "<new-name1 id="1">recU</new-name1>" and "recU1" (recU :cat) and "<new-name2 id="2">recS</new-name2>" 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. "

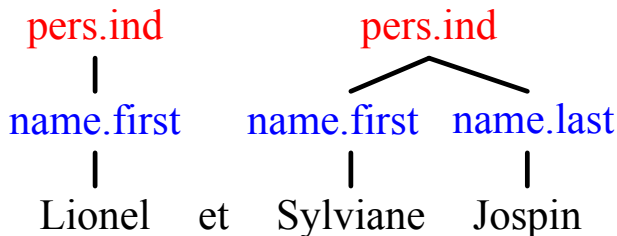
Renommage de gènes

		A1			
		Former	New	Nothing	Total
A2	Former	71	13	23	107
	New	8	69	15	92
	Nothing	7	8	18 840	18 855
	Total	86	90	18 878	19 054

Changeons de point de vue

		A1			
		Former	New	Nothing	Gene names
A2	Former	71	13	23	107
	New	8	69	15	92
	Nothing	7	8	951	966
	Gene names	86	90	989	1 165

Entités nommées structurées [LAW 2011]



Les “annotables”

Annotables	Annotateurs	Experts
U1 : n-grams	$\kappa = 0,84522$ $\pi = 0,81687$	$\kappa = 0,91123$ $\pi = 0,90258$
U2 : n-grams ≤ 6	$\kappa = 0,84519$ $\pi = 0,81685$	$\kappa = 0,91121$ $\pi = 0,90257$
U3 : NPs	$\kappa = 0,84458$ $\pi = 0,81628$	$\kappa = 0,91084$ $\pi = 0,90219$
U4 : entités Ester	$\kappa = 0,71300$ $\pi = 0,71210$	$\kappa = 0,82607$ $\pi = 0,82598$

Accords inter-annotateurs de la campagne d'annotation en entités nommées structurées

(κ correspond au Kappa de Cohen et π au Pi de Scott)

Exercice : algorithme d'estimation des annotables

3 groupes

Trouver un algorithme permettant d'approcher au mieux le nombre d'annotables

Les "annotables" : proposition

Algorithme EM

Entrée : $\{M_j\}$ (ensemble des éléments annotés par les annotateurs A_j)

$$\delta_0 = \left| \bigcup_j M_j \right|$$

for (i=1; change in δ ; i++) **do**

$$\text{expectation : } \gamma_i = \prod_j P(A_j \text{ manque un annotable}) = \prod_j \frac{\delta_{i-1} - |M_j|}{\delta_{i-1}}$$

$$\text{maximization : } \delta_i = \frac{\delta_0}{1 - \gamma_i}$$

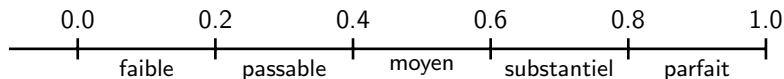
end for

return δ

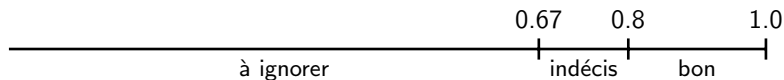
- 1 Au cœur des systèmes, les ressources langagières
- 2 Focus sur le Crowdsourcing
- 3 **Accord(s) inter-annotateurs**
 - Exercice
 - Prendre en compte le hasard
 - Coefficient S
 - Coefficient π
 - Coefficient κ
 - Kappa et annotables
 - **Interprétation**
- 4 Conclusion

Évolutions de l'interprétation des Kappas

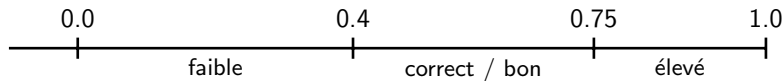
[Landis and Koch, 1977]



[Krippendorff, 1980]



[Green, 1997]



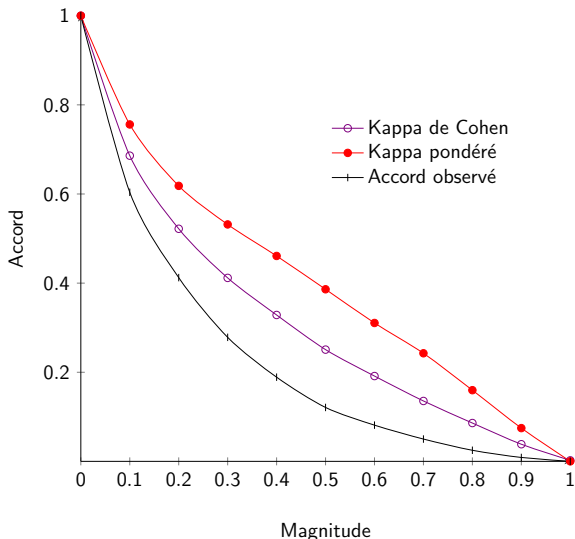
Donner du sens au résultat obtenu [COLING 2012a]

Création d'un outil "Richter" qui :

- prend en entrée une annotation de référence (réelle ou générée automatiquement)
- génère des annotations dégradées en fonction d'une **magnitude** (de 0 à 1)
- applique une ou des mesures d'accord inter-annotateurs sur chaque ensemble d'annotations (correspondant à une magnitude)

Sur le corpus TCOF-POS [TALN 2012]

Pas de prévalence, mais proximité entre catégories prise en compte :



Origines du cours

Extérieures :

- Cours d'Alain Couillault, Gestion sémantique des contenus, Master ICONE, La Rochelle.
- http://faculty.washington.edu/ebender/2012_472/0404.pdf
- <http://www.cis.upenn.edu/~cis639/docs/fsexamples.html>

Personnelles :

- Cours de Master Sciences Cognitives option TAL
<http://www.loria.fr/~fortkare/Enseignement.php>
- Séminaire ICAR
<http://icar.univ-lyon2.fr/pages/seminaires.htm>



Adda, G., Sagot, B., Fort, K., and Mariani, J. (2011).

Crowdsourcing for language resource development : Critical analysis of Amazon Mechanical Turk overpowering use.

In Proceedings of the Language and Technology Conference (LTC), Poznań, Pologne.

5 pages.



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.



Artstein, R. and Poesio, M. (2008).

Inter-coder agreement for computational linguistics.

Computational Linguistics, 34(4) :555–596.



Bazillon, T., Estève, Y., and Luzzati, D. (2008).

Transcription manuelle vs assistée de la parole préparée et spontanée.
In Actes de Journées d'Études sur la Parole (JEP), Avignon, France.



Bennett, E. M., Alpert, R., and C. Goldstein, A. (1954).
Communications through limited questioning.
Public Opinion Quarterly, 18(3) :303–308.



Benzitoun, C., Fort, K., and Sagot, B. (2012).
TCOF-POS : un corpus libre de français parlé annoté en
morphosyntaxe.
In Actes de Traitement Automatique des Langues Naturelles (TALN),
pages 99–112, Grenoble, France.



Böhmová, A., Hajič, J., Hajičová, E., and Hladká, B. (2001).
The prague dependency treebank : Three-level annotation scenario.
In Abeillé, A., editor, Treebanks : Building and Using Syntactically
Annotated Corpora. Kluwer Academic Publishers.



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., Merklings, 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).



Chamberlain, J., Fort, K., Kruschwitz, U., Lafourcade, M., and Poesio, M. (2012).

Theory and Applications of Natural Language Processing, chapter Using Games to Create Language Resources : Successes and Limitations of the Approach.

Theory and Applications of Natural Language Processing. Springer.
A paraître.



Chamberlain, J., Poesio, M., and Kruschwitz, U. (2008).

Phrase Detectives : a web-based collaborative annotation game.

In Proceedings of the International Conference on Semantic Systems (I-Semantics'08), Graz, Autriche.



Cohen, J. (1960).

A coefficient of agreement for nominal scales.

Educational and Psychological Measurement, 20(1) :37–46.



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, Singapur.



Deneme (2009).

How many turkers are there ?

[http ://groups.csail.mit.edu/uid/deneme/](http://groups.csail.mit.edu/uid/deneme/).



Fort, K., Adda, G., and Cohen, K. B. (2011).

Amazon Mechanical Turk : Gold mine or coal mine ?

Computational Linguistics (editorial), 37(2) :413–420.



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

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. 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.



Green, A. M. (1997).

Kappa statistics for multiple raters using categorical classifications.

In Proceedings of the Twenty-Second Annual Conference of SAS Users Group, San Diego, USA.



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.



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.



Kim, J.-D., Ohta, T., and Tsujii, J. (2008).

Corpus annotation for mining biomedical events from literature.

BMC Bioinformatics, 9(1) :10.



Krippendorff, K. (1980).

Content Analysis : An Introduction to Its Methodology, chapter 12.

Sage, Beverly Hills, CA., USA.



Lafourcade, M. (2007).

Making people play for lexical acquisition.

In Proceedings of the 7th Symposium on Natural Language Processing (SNLP 2007), Pattaya, Thaïlande.



Landis, J. R. and Koch, G. G. (1977).

The measurement of observer agreement for categorical data.

Biometrics, 33(1) :159–174.



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.



Mathet, Y., Widlöcher, A., Fort, K., François, C., Galibert, O., Grouin, C., Kahn, J., Rosset, S., and Zweigenbaum, P. (2012).

Manual corpus annotation : Evaluating the evaluation metrics.

In Proceedings of the International Conference on Computational Linguistics (COLING), pages 809–818, Mumbai, Inde.

Poster.



Scott, W. A. (1955).

Reliability of content analysis : The case of nominal scale coding.

Public Opinion Quarterly, 19(3) :321–325.



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.



von Ahn, L. (2006).

Games with a purpose.

IEEE Computer Magazine, pages 96–98.



Voormann, H. and Gut, U. (2008).

Agile corpus creation.

Corpus Linguistics and Linguistic Theory, 4(2) :235–251.