Corpus
Annotation

Béatrice Daille

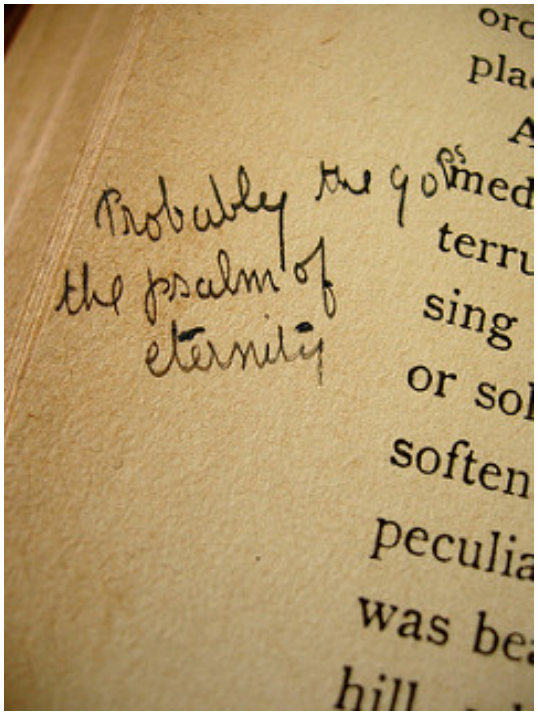
Faculté des Sciences et Techniques de Nantes
Département Informatique

2018-2019

Corpus

- Corpus : exemples, définition, caractérisation
- Métadonnées : Dublin Core, TEI
- Web comme corpus
- **Annotation des corpus**
- Exploration des corpus : phrasémes, collocations

Annotation : définition



L'annotation recouvre à la fois le processus consistant à apposer (ad-) une note sur un support, l'ensemble des notes ou chaque note particulière qui en résulte et ce, sans préjuger a priori de la nature du support considéré (texte, vidéo, images, etc.), du contenu sémantique de la note (note chiffrée, valeur choisie dans un référentiel fermé ou texte libre), de son positionnement global ou local, ni de son objectif (visée évaluative ou caractérisante, simple commentaire discursif).

Objectifs

Utilité des annotations

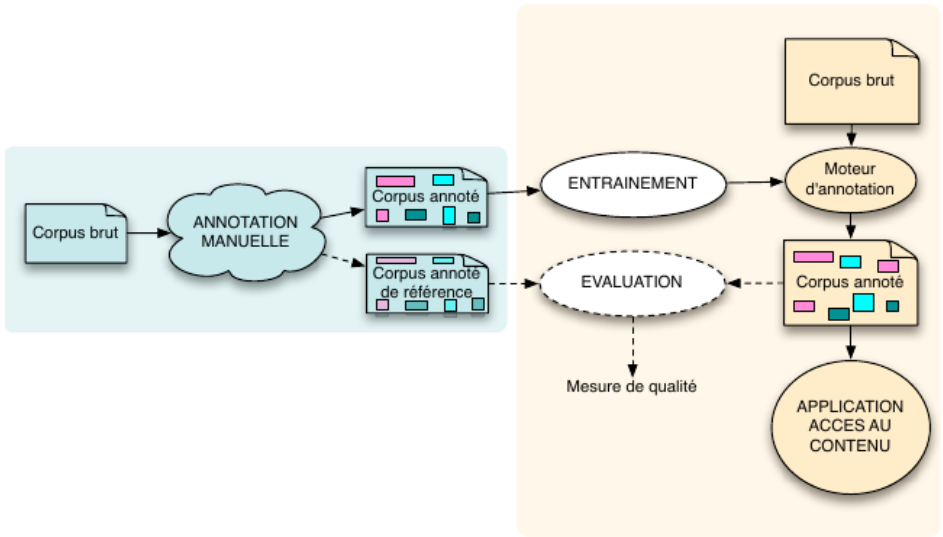
TALN

1. Acquisition : classifieur construit par apprentissage supervisé
2. Évaluation : annotation de référence pour évaluer un algorithme

Linguistique

1. Exploration d'un corpus
2. Étude d'un phénomène linguistique à grande échelle

Annotation pour le TAL



Architecture schéma d'annotation

Identique à l'architecture d'un SGBD

1. Niveau externe : visualisé et manipulé par les utilisateurs
2. Niveau intermédiaire : schéma logique (spécification des annotations)
3. Niveau physique : codage et stockage des annotations (annotations déportées ou insérées, XML et formats linéaires)

Des ressources diverses

Diversité

- médias : texte, parole, musique, vidéo
- champs d'applications : linguistique, domaines spécialisés

Complexité

- 1960-1990 : morpho-syntaxe
- 1990-2000 : morpho-syntaxe et syntaxe (corpus arborés), entités nommées simples, sémantique (WordNet, FrameNet)
- depuis 2000 : annotations sémantiques variées (opinions, émotions, etc.), entités nommées structurées, etc.

Taille

- Brown Corpus, 1 million de mots
- BNC, 100 millions de mots

Types d'annotation

Extrême diversité des annotations

- niveau document/texte ou signal
- expressivité (booléen, catégories, relations)
- complexité : phénomènes linguistiques complexes, ontologies

Annotations au niveau du document

Intégration dans les méta-données
Profilage du texte : genre, discours, positionnement, niveau de langue, domaine

Annotations au niveau du texte

Enrichissements linguistiques :

- Segmentation : phrasèmes, entités nommées
- Catégories grammaticales
- Lemmatisation
- Analyse syntaxique

Catégories grammaticales universelles

Jeu de catégories grammaticales validées sur 22 langues
Tables de correspondance pour 27 jeux d'étiquettes

VERB	verbes (tous les temps et modes)
NOUN	noms (commun et propre)
PRON	pronoms
ADJ	adjectifs et numéraux ordinaux
ADV	adverbes
ADP	adpositions (prépositions and postpositions)
CONJ	conjonctions
DET	déterminants (regroupent les articles et adjectifs possessifs, démonstratifs, interrogatifs, exclamatifs ou numéraux cardinaux)
NUM	chiffres
PRT	particules ou autres mots fonctionnels
X	autres : mots étrangers, typos, abréviations, interjections
.	ponctuation

`http://code.google.com/p/universal-pos-tags/`
"A Universal Part-of-Speech Tagset" by Slav Petrov, Dipanjan Das and Ryan McDonald LREC 2012

Exemple annotation catégories grammaticales

Text				
Cyndi savoredthe soup.				
^0...^5...^10..^15..^20				
Annotations				
Id	Type	SpanStart	Span End	Features
1	token	0	5	pos=NP
2	token	6	13	pos=VBD
3	token	14	17	pos=DT
4	token	18	22	pos=NN
5	token	22	23	
6	name	0	5	name_type=person
7	sentence	0	23	constituents=[1],[2],[3].[4],[5]

Corpus arborés

- Grammaires de constituants
- simple parenthésage des constituants
  [Ce guide][[leur] permet [de [se familiariser [. . .]
- étiquetage des constituants
  [N Ce_DEDEMMS guide_NCOMS N][
  [P leur_PPCA6MP P] permet_V_VINIP3
  [P de_PREPD [Vi se_PPRE6MP familiariser_VPRN [. . .]

Corpus arborés

- Grammaires de dépendances
- indication des relations de dépendances
  Ce @DN>
  guide @NV2>
  leur @PV>
  permet
  [...]
- Corpus Suzanne

Corpus arborés

- Grammaires de constituants et grammaires de dépendances
- indication des relations fonctionnelles
  [N <Sujet> Ce_DEDEMMS guide_NCOMS N]
  [V [P <ObjetIndirect> leur_PPCA6MP P] permet V_VINIP3 [. . .]
- relations logiques ou profondes
  [N Ce_DEDEMMS guide_NCOMS N]
  [V [P <8> leur_PPCA6MP P] permet V_VINIP3
  [P de_PREPD [N <8> N] [Vi se_PPRE6MP familiariser_VPRN [. . .]
- Pen Treebank

Corpus arborés : Penn Treebank

- 4,8 M de tokens en anglais américain (Brown corpus ré-annoté)
- annotation en morpho-syntaxe : 48 catégories (12 ponctuations et symboles)
- annotation en syntaxe : 15 catégories

Dépendances universelles

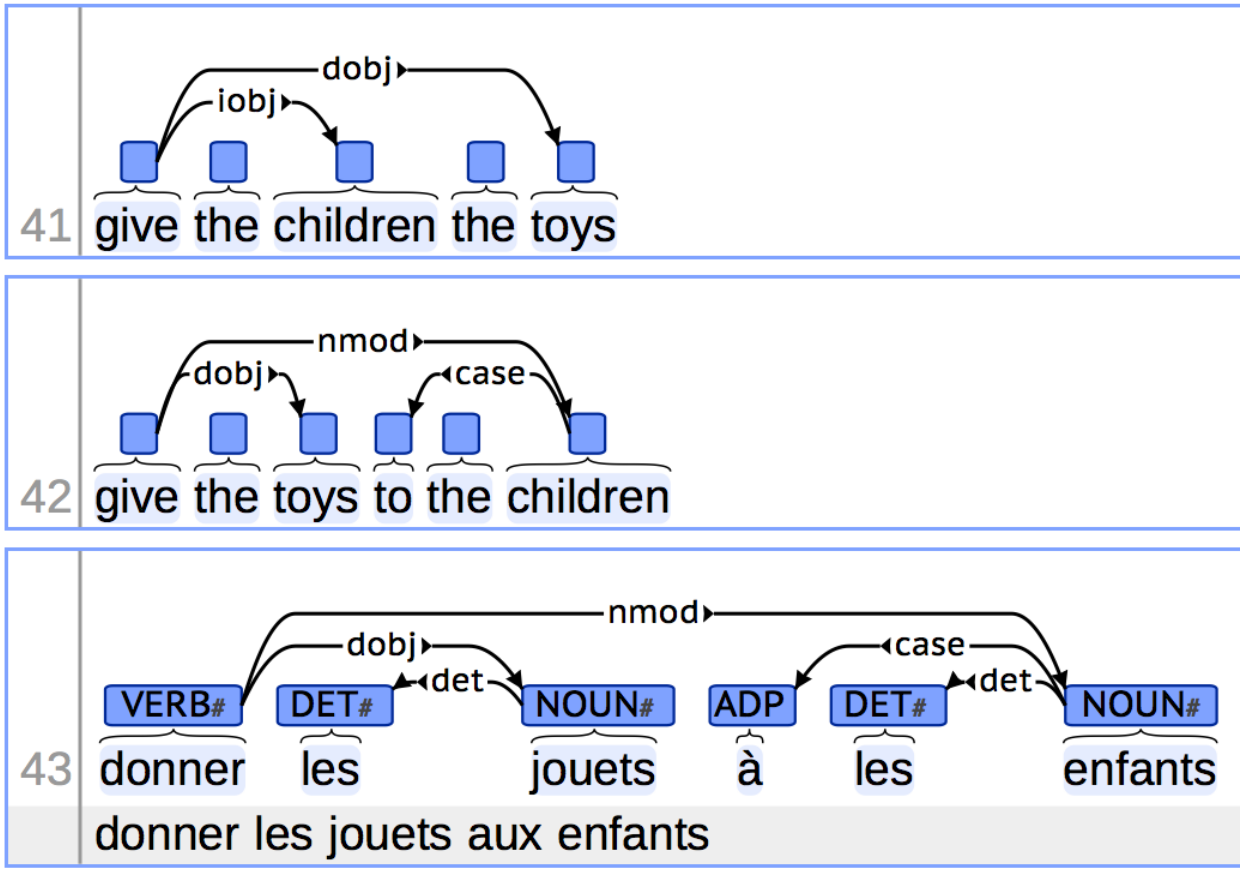
37 dépendances élémentaires

Clausal Argument Relations	Description
NSUBJ	Nominal subject
DOBJ	Direct object
IOBJ	Indirect object
CCOMP	Clausal complement
XCOMP	Open clausal complement
Nominal Modifier Relations	**Description**
NMOD	Nominal modifier
AMOD	Adjectival modifier
NUMMOD	Numeric modifier
APPOS	Appositional modifier
DET	Determiner
CASE	Prepositions, postpositions and other case markers
Other Notable Relations	**Description**
CONJ	Conjunct
CC	Coordinating conjunction

Figure 14.2 Selected dependency relations from the Universal Dependency set. (de Marneffe et al., 2014)

`http://universaldependencies.org/u/dep/index.html`
"Universal Stanford Dependencies : A cross-linguistic typology" by Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, Christopher D. Manning. LREC 2014

Exemples


41
dobj
iobj
give the children the toys
42
nmod
dobj
case
give the toys to the children
43
nmod
dobj
det
case
det
VERB#
DET#
NOUN#
ADP
DET#
NOUN#
donner les jouets à les enfants
donner les jouets aux enfants

Annotations au niveau du texte

Phénomènes linguistiques complexes :

- Lexèmes, termes
- Anaphores
- Dialogue
- Expressions d'évaluations

ISO-TimeML

Expressions temporelles

les événements dans le temps présents dans les articles de presse, les biographies, Wikipédia, etc.

TIMEX3 attribut TYPE

DATE	le 15 janvier 2019, hier, l'été 2006
TIME	11h30, treize heures quinze, la nuit dernière
DURATION	2 mois, trois jours,
SET	deux fois par mois, tous les 2 jours

Tout l'empan de l'expression référentielle est marqué.

ISO-TimeML

Les événements

la notion d'événement recouvre tous les types de situations (états, activités, achèvements, etc.) EVENT attribut CLASS

OCCURRENCE vendre, construire, tuer

STATE amoureux, fatigué

REPORTING dire, annoncer

I-Action essayer, promettre

I-State croire, vouloir

Aspectual commencer, terminer, arrêter

Perception voir, entendre, sentir

autres attributs : TYPE, TENSE, ASPECT, POLARITY, MODALITY
balise sur la tête du groupe événementiel, en excluant les auxiliaires, les modifieurs, les adverbes de négation, les clitiques, etc.

ISO-TimeML

Les relations

TLINK relation temporelle entre un événement et une date.

ALINK relation aspectuelle. Par exemple, entre un verbe aspectuel (commencer, cesser, continuer) et son complément événementiel.

SLINK relation de subordination. Par exemple, la relation qui existe entre un verbe modal (falloir, devoir) ou de perception (voir, entendre) et son complément événementiel.

MLINK relation pour mesurer la durée d'un évènement

attribut RELTYPE pour spécifier la relation :
TLINK : BEFORE, IS_INCLUDED

ISO-TimeML

Les marqueurs de temporalité

SIGNAL

prépositions temporelles sur, dans, avant

conjonction de subordination temporelles avant

caractères spéciaux - et /

le marqueur est identifié lors d'une relation TLINK

Exemple

Jean est <EVENT id="e1" class="OCCURRENCE" pos="VERB" tense="PAST" vform="PASTPART"> *né* </EVENT>
<SIGNAL id="s1"> *avant* < /SIGNAL>
l' <EVENT id="e2" class="OCCURRENCE" pos="NOUN"> *introduction* < /EVENT> *de l'euro*.
<TLINK id="l1" eventID="e1" relatedToEvent="e2" signalID="s1" relType="BEFORE" / >

Annotations au niveau du texte

Correction

- Normalisation : détection des erreurs, formes non standards
- Anonymisation

Coût annotation

Penn Tree Bank (PTB)

PTB1 (Marcus et al. 1993)

- correction de l'étiquetage morpho-syntaxique : 3 000 mots à l'heure, 3 H par jour
- correction de l'étiquetage syntaxique : 750 mots à l'heure, 3 H par jour
- + courbe d'apprentissage de 1 mois (étiquettes morpho-syntaxiques) à 2 mois (syntaxe)

Coût annotation

Prague Dependancy Treebank

- 1996-2004 (Böhmová et al. 2001)
- construit à partir du CNC (Czech National Corpus)
- 3 niveaux de structures
  1. morphologique : 1,8 millions de mots
  2. analytique : syntaxe en dépendances
  3. sémantique

Coût annotation

Prague Dependancy Treebank

Version 1.0

- annotation manuelle des niveaux morphologique et analytique
- temps : 5 ans
- nombre de personnes : 22 dont 17 simultanément pendant certaines périodes
- coût : 600 000 $

Coût annotation

GENIA

- (Kim et al. 2008)
- 400 000 mots annotés en microbiologie
- 5 annotateurs à mi-temps, 1 coordinateur senior 1 coordinateur junior pendant 1,5 an

Coût annotation

ESTER

- 100 heures de parole transcrite (campagne d'évaluation ESTER, systèmes de transcription, 2008)
- 1 h de parole = entre 20 à 60 H de transcription

Consensus au cœur de l'annotation

- guide d'annotation
- réunion avec les annotateurs, plusieurs phases : mise au point, vérification de l'applicabilité du guide d'annotation
- mesurer le consensus, garantie d'une cohérence

Dimensions de complexité de l'annotation

(Fort et al. 2012)

1. discrimination des unités à annoter
2. délimitation des unités à annoter
3. expressivité du langage d'annotation
4. jeu d'étiquettes
5. ambiguïté
6. contexte textuel à consulter, connaissances à mobiliser

Les guides d'annotation

Des ressources précieuses

Limiter le coût de l'annotation

- outils d'annotation : WebAnno, Glozz, GATE, Knowtator, Callisto, Prat, etc. $\rightarrow$ pas toujours adaptés
- propagation d'étiquettes, pré-annotation, apprentissage actif
- myriadisation : Amazon Mechanical Turk et jeux ayant un but

Mesures d'accord entre annotateurs

- Mesures d'accord entre deux annotateurs
  - accord observé A_0
  - accord attendu A_e
    - S de Scott A_e^S
    - π de Scott A_e^π
    - κ de Cohen A_e^κ
- Interprétation de la valeur des accords
- Accord inter-annotateurs personnalisé
- Analyse de l'accord inter-annotateur global

Accord observé

Pourcentage d'accord ou accord observé A_o

$$A_o = \frac{1}{\mathbf{i}} \sum_{i \in I} \text{agr}_i \tag{1}$$

avec **i** : nombre total d'éléments annotables
et agr : accord sur l'annotation

Limitation :

- Pas de prise en compte du hasard : un petit nombre de catégories donne un meilleur accord observé
- Pas de compensation en fonction de la distribution des éléments dans les catégories : une catégorie prépondérante influence largement l'accord observé

Exemple accord observé

Annotation du registre de langue
Matrice de confusion : 2 annotateurs : Paul et Marie, 3 catégories : Familier, Courant, Soutenu, 3 fichiers

		Paul	Paul	Paul	
		F	C	S	Somme
Marie	F	0	0	0	0
Marie	C	1	2	0	3
Marie	S	0	0	0	0
Somme		1	2	0	3

$A_o = \frac{1}{3}.2 = 0{,}66$

Accord attendu

Prise en compte que les annotateurs classent un élément quelconque dans une catégorie au hasard A_e

$$S\pi\kappa = \frac{A_o - A_e}{1 - A_e} \tag{2}$$

S : annotations réalisées au hasard suivent une distribution uniforme dans les catégories (Bennett et al. 1954)

$$A_o^S = \frac{1}{\mathbf{k}} \tag{3}$$

k : nombre de catégories k

Calcul sur exemple précédent : $A_o^S = \frac{1}{3} = 0{,}33$ $\qquad$ $S = \frac{0{,}66 - 0{,}33}{1 - 0{,}33} = 0{,}49$

- Problème : un petit nombre de catégories donne un meilleur accord observé

π de Scott

Accord attendu en supposant que la répartition des éléments dans les catégories n'est pas homogène A_e^π (Scott, 1955)
Estimation de la répartition moyenne réalisée par les annotateurs

$$A_e^\pi = \sum_{k \in K} (\frac{\mathbf{n}_k}{2\mathbf{i}})^2 \tag{4}$$

$$\pi = \frac{A_o - A_e^\pi}{1 - A_e^\pi} \tag{5}$$

avec $\mathbf{n}_k$: nombre d'affectations à k par les deux annotateurs
$\mathbf{i}$: nombre d'éléments annotables i
k : une catégorie

Exemple calcul Scott

		Paul	Paul	Paul	
		F	C	S	
Marie	F	a	b	c	C^F_{Marie}
Marie	C	d	e	f	C^C_{Marie}
Marie	S	g	h	j	C^S_{Marie}
		C^F_{Paul}	C^C_{Paul}	C^S_{Paul}	

		P	P	P	
		F	C	S	
M	F	0	0	0	0
M	C	1	2	0	3
M	S	0	0	0	0
		1	2	0	

C^F_{Marie}= a+b+c

$\mathbf{n}_F = C^F_{Marie} + C^F_{Paul}$

$$A^{\pi}_{e} = \frac{(\frac{\mathbf{n}_F}{2})^2+(\frac{\mathbf{n}_C}{2})^2+(\frac{\mathbf{n}_S}{2})^2}{\mathbf{i}^2} = \frac{(\frac{1+0}{2})^2+(\frac{3+2}{2})^2+0}{9} = 0{,}72$$

$$\pi = \frac{0{,}66-0{,}72}{1-0{,}72} = \text{-}0{,}23$$

κ de Cohen

Modélisation du hasard A_o^κ (Cohen 1960)
Répartition des éléments entre catégories peut être différent pour chaque annotateur

$$A_e^\kappa = \sum_{k \in K} \frac{\mathbf{n}_{c1k}}{\mathbf{i}} . \frac{\mathbf{n}_{c2k}}{\mathbf{i}} \tag{6}$$

$$\kappa = \frac{A_o - A_e^\kappa}{1 - A_e^\kappa} \tag{7}$$

avec $\mathbf{n}_{c1k}$: nombre d'affectations à k par l'annotateur $c1$
$\mathbf{i}$: nombre d'éléments annotables i

Exemple calcul Cohen

		Paul	Paul	Paul
		F	C	S
Marie	F	a	b	c
Marie	C	d	e	f
Marie	S	g	h	j

		Paul	Paul	Paul
		F	C	S
Marie	F	0	0	0
Marie	C	1	2	0
Marie	S	0	0	0

$$A_o = \frac{a+e+i}{a+b+...+j} = \frac{2}{3} = 0{,}66$$

$$\mathbf{n}_{c1F} = \frac{C^F_{Marie}}{i} = \frac{a+b+c}{a+b+...+j} = \frac{0}{3} = 0 \qquad \mathbf{n}_{c2F} = \frac{C^F_{Paul}}{i} = \frac{a+d+g}{a+b+...+j} = \frac{1}{3} = 0{,}33$$

$$\mathbf{n}_{c1C} = \frac{C^C_{Marie}}{i} = \frac{d+e+f}{a+b+...+j} = \frac{3}{3} = 1 \qquad \mathbf{n}_{c2C} = \frac{C^C_{Paul}}{i} = \frac{b+e+h}{a+b+...+j} = \frac{2}{3} = 0{,}66$$

$$A^{\kappa}_e = \mathbf{n}_{c1F}.\mathbf{n}_{c2F} + \mathbf{n}_{c1C}.\mathbf{n}_{c2C} + \mathbf{n}_{c1S}.\mathbf{n}_{c2S} = 0 + 0{,}66 + 0 = 0{,}66$$

$$\kappa = \frac{0{,}66-0{,}66}{1-0{,}66} = 0$$

Interprétation

Valeur prise dans [-$\frac{A_e}{1-A_e}$, 1] avec 1 accord total

- Ordre $S \geq \pi$ et $\pi \leq \kappa$
- si $\pi < \kappa$ et π et κ proches : peu de biais entre les annotateurs

$0 \leq 0{,}2$	faible	Landis et Koch (1977)
$0{,}2 \leq 0{,}4$	passable	
$0{,}4 \leq 0{,}6$	moyen	
$0{,}6 \leq 0{,}8$	substantiel	
$0{,}8 \leq 0{,}1$	parfait	
$0 \leq 0{,}67$	à ignorer	Krippendorff (1980)
$0{,}67 \leq 0{,}8$	indécis	
$0{,}8 \leq 1$	bon	
$0 \leq 0{,}4$	faible	Green (1997)
$0{,}4 \leq 0{,}75$	correct/bon	
$0{,}75 \leq 1$	élevé	

Interprétation

S : ne dépend que du nombre de catégories
$\rightarrow$ n'est pas sensible à la répartition des éléments dans les catégories

π et κ : lorsque les catégories sont fortement disproportionnées, et même si fort accord sur la catégorie prédominante, très sensibles aux désaccords sur les catégories minoritaires

multi π ou Fleiss-κ

généralisation de π de Scott à plusieurs annotateurs
le nombre de paires d'annotateurs en accord par rapport à toutes les paires de jugements possibles pour l'élément considéré

$$\text{agr}_i = \frac{1}{\mathbf{c}(\mathbf{c}-1)} \sum_{k \in K} \mathbf{n}_{ik}(\mathbf{n}_{ik}-1) \tag{8}$$

$$A_o^{\pi-m} = \frac{1}{\mathbf{i}} \sum_{i \in I} \text{agr}_i \tag{9}$$

$$A_e^{\pi-m} = \frac{1}{(\mathbf{ic})^2} \sum_{k \in K} (\mathbf{n}_k)^2 \tag{10}$$

$$\pi - m = \frac{A_o^{\pi-m} - A_e^{\pi-m}}{1 - A_e^{\pi-m}} \tag{11}$$

$\mathbf{c}$: nombre d'annotateurs c

Exemple de calcul de Fleiss-κ

4 annotateurs, 3 catégories, 3 fichiers

n_{ik}	F	C	S	P_c
Fic1	0	4	0	$P_{Fic1} = \frac{1}{4.3}(0 + 4 * 3 + 0)$
Fic2	1	3	0	$P_{Fic2} = \frac{1}{4.3}(1 * 0 + 3 * 2 + 0)$
Fic3	0	3	1	$P_{Fic3} = \frac{1}{4.3}(0 + 3 * 2 + 0)$
Somme	1	10	1	12
P_k	$P_F = \frac{1}{12}$ 0,08	$P_C = \frac{10}{12}$ 0,83	$P_S = \frac{1}{12}$ 0,08	

$\sum_{i \in I} P_i$= 1+0,5+0,5=2

$A_o^{\pi - m} = \frac{1}{3}.2$=0,67

$A_e^{\pi - m}$=$\frac{1^2+10^2+1^2}{(3.4)^2} = \frac{112}{144}$= 0,77

$\pi - m$=$\frac{0,67-0,77}{1-0,77}$= -0,30

Hypothèses

- Les annotateurs annotent tous les éléments $\rightarrow$ pas toujours le cas en situation réelle
- La distance entre les catégories est identique $\rightarrow$ des classes peuvent être plus proches que d'autres
  possibilité de pondérer les catégories : indice α (Krippendorff, 2004)
  définition d'une matrice de pondération adaptée aux catégories
  par exemple : pour l'annotation du registre, un désaccord 0 entre familier et soutenu ; 0,5 entre courant et soutenu ou courant et familier
  $A_e^{\pi-pm} = \frac{1}{(\mathbf{ic})^2} \sum_{k \in K} \sum_{k' \in K} \text{pond}(k, k') \mathbf{n}_k \mathbf{n}'_k$
- Tous les annotateurs n'ont pas la même expertise $\rightarrow$ comparaison des accords entre paires d'annotateurs et éventuellement sur certaines catégories

Références

La bible
Handbook of Linguistic Annotation. Ide and Pustejovsky (eds). Springer 2017.

Mesures d'accords
Ron Artstein et Massimo Poesio : Inter-coder agreement for computational linguistics. Computational Linguistics , 34(4) :555–596, 2008. ISSN 0891-2017.