

➤ *Annotation de corpus*

- une « valeur ajoutée », consistant en un apport d'informations de nature *interprétative* aux données brutes

Leech (1997:2-4)

Pourquoi annoter

- indexer et archiver le corpus
- consulter le corpus, y faire des recherches
- extraire de l'information

De nombreuses tâches peuvent se formuler comme des tâches d'annotation

- annotation sert à la segmentation car elle délimite des fragments de données ;
- annotation regroupe des segments pour leur affecter une catégorie ;
- annotation met en relation des fragments.

Habert (2005)

- transcription est un enrichissement de l'information sonore au moyen d'une information orthographique

Unités déjà définies : Les unités sont déjà délimitées, l'annotateur doit alors les caractériser.

Ancrage minimal : L'annotateur marque une position dans le flux textuel.

Segmentation : L'annotateur doit pavier le texte.

Unitizing : L'annotateur doit repérer dans le texte les unités ; il peut aussi caractériser ces mêmes unités.

Mise en relation : L'annotateur doit relier deux (ou plus) unités entre elles.

Diagram illustrating 'Unités déjà définies'. It shows the word 'Lorem' followed by five words ('Ipsum', 'dolor', 'sit', 'amet') each preceded by a red letter (A or B). The letters are positioned at the start of their respective words.

(a) Unités déjà définies

Diagram illustrating 'Position minimale'. It shows the first part of a Latin sentence: 'Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum nec libero at arcu sodales semper. Suspendisse in porta eros. Cras a dolor tincidunt, volutpat lorem id, ultricies enim.' Two red circles with numbers 1 and 2 are placed above the text, indicating specific positions.

(b) Position minimale

Diagram illustrating 'Segmentation'. It shows the same sentence as in (b), but with vertical red lines (punctuation marks) placed after each word, effectively dividing the sentence into individual words.

(c) Segmentation

Diagram illustrating 'Unitizing'. It shows the sentence with several words highlighted in red boxes: 'ipsum', 'adipiscing', 'elit.', 'porta', 'eros.', 'dolor', 'tincidunt', 'lorem', 'id', 'ultricies', 'enim.', 'justo', 'lobortis', 'elementum', 'vitae', 'at', 'ligula.', 'tristique', 'senectus', 'netus', 'malesuada', 'fames', 'turpis', 'egestas.'

(d) Unitizing

Diagram illustrating 'Mise en relation'. It shows the sentence with multiple words highlighted in red boxes: 'ipsum', 'adipiscing', 'elit.', 'porta', 'eros.', 'dolor', 'tincidunt', 'lorem', 'id', 'ultricies', 'enim.', 'justo', 'lobortis', 'elementum', 'vitae', 'at', 'ligula.', 'tristique', 'senectus', 'netus', 'malesuada', 'fames', 'turpis', 'egestas.'. The highlighted words correspond to the ones in (d).

(e) Mise en relation



➤ annotation intégrale

- chaque mot fait l'objet de l'étiquetage

➤ annotation partielle

- les renseignements attachés à certains mots sont inexistant ou incomplets



- Une étiquette
- Multi-étiquette
 - Assigner plusieurs étiquettes à une même unité
 - Plusieurs émotions, sentiments

Types de catégories

➤ catégories binaires

- présence ou non du phénomène, positif ou négatif, etc.,

➤ nominales

- spectre de catégories plus large

➤ scalaires

- échelle de valeur

➤ certitude

- incertains ou ambigus

Structuration d'étiquettes

- Catégories / sous-catégories
 - DET / DETDEM+DETPOS+...
- Catégories / attributs
 - <DET type="dem"/>

Contenu de l'annotation

- Métadonnées
- Segmentation
- Annotation syntaxique
- Annotation sémantique

Outils d'annotation

➤ traitements de texte

- OpenOffice Writer, Microsoft Word, Emacs, éditeurs XML (Oxygen, XMLmind), Notepad++, SublimeText, etc.

➤ tableurs

The screenshot shows a WYSIWYG text editor window with a toolbar at the top, a main text area in the center, and a styles panel on the right.

Toolbar: Includes standard file operations (Fichier, Édition, Affichage, Insertion, Format, Styles, Tableau, Formulaire, Outils, Fenêtre, Aide), font selection (Style par défaut, Liberation Se), size (12), and various styling icons (bold, italic, underline, superscript, subscript, etc.).

Text Area: Contains the following text:

Aramis, après un voyage en Lorraine, disparut tout à coup et cessa d'écrire à ses amis. On apprit plus tard, par Mme de Chevreuse, qui le dit à deux ou trois de ses amants, que cédant à sa vocation, il s'était retiré dans un couvent ; seulement on ne sut jamais lequel.
Bazin devint frère laï.
Athos resta mousquetaire sous les ordres de d'Artagnan jusqu'en 1633, époque à laquelle, en revenant d'un voyage qu'il fit en Roussillon, il quitta aussi le service, sous prétexte qu'il venait de recueillir un petit héritage dans le Blaisois.
Grimaud suivit Athos.
D'Artagnan se battit trois fois avec Rochefort et le blessa trois fois.

Styles Panel: Shows a list of styles with EN.Personne selected (highlighted in orange). Other styles listed are EN.Date and EN.Lieu. There is also a "Style par défaut" entry. At the bottom of the panel are checkboxes for "Afficher les aperçus" (Preview) and "Hiérarchie" (Hierarchy).

FIGURE 1.5 – Exemple d'annotation via l'interface d'un traitement de textes WYSIWYG. Texte provenant de l'épilogue des « Trois mousquetaires » d'Alexandre Dumas.

	A	B
1	Entités nommées	Catégorie
2	Aramis	Personne
3	Lorraine	Lieu
4	Mme de Chevreuse	Personne
5	Bazin	Personne
6	Athos	Personne
7	D'Artagnan	Personne
8	1633	Date
9	Roussillon	Lieu
10	Blaisois	Lieu
11	Grimaud	Personne
12	Rochefort	Personne

FIGURE 1.6 – Exemple d’entités nommées annotées via un tableur.

Outils d'annotation manuelle

➤ Interfaces web

- on peut les installer sur un serveur et rendre accessibles via un site Web.
- L'interface ressemble alors à un logiciel d'annotation *standalone*, avec les mêmes possibilités de fonctionnalités proposées.
 - l'annotateur n'a pas à installer directement un logiciel
 - le travail collaboratif

- Brat

➤ Outils à installer : Glozz, Gate, Analec

- Inception

<https://inception-project.github.io/>

Anelec 1.0 : Echenoz7 - vue : Echenoz?



Documents Structure Vue Texte Unités Relations Schémas Règles Statistiques

Gestion des unités

Unité : Forme explicite Forme explicite 35 0 Crée Supprimer Rectifier une borne

... vieilles reconstructions qui se collaient en grincant contre lui, terrifiées par le plan d'acquisition des sols. En manque de marchandise, son portefeuille de maillures portait le nom (Wagner) de l'architecte-sculpteur gravé dans un cartouche en haut à crois. Et le mur sur lequel, avec toute son équipe, l'artiste Flers avait peint pour figurer Sylvie Fabre en pied, surplombait un petit espace vert rudimentaire, sorte de square sans accessoires qui ne consistait qu'à former le coin de la rue.

Choisis par Flers, pressée par Fabre, Sylvie avait accepté de poser. Elle n'avait pas aimé cela. C'était trois ans avant la naissance de Paul, pour qui ce mur n'avait qu'une tranche de vie antérieure. Regarder un peu sa mère, s'émerveillait Fabre que ce spectacle n'était en larmes, en rire, selon. Mais il pouvait aussi chercher la scène, se faire franchement hostile à l'effigie contre laquelle, en écho, rebondissaient ses reproches – Paul s'occupait de modérer le plus dès qu'un strouppement menaçait de se former.

Plus tard, suffisamment séparé de Fabre pour qu'on ne se parlât même plus, Paul visitait sa mère sur un rythme plus coupé, deux ou trois fois par mois,

Champs de l'unité : "Forme explicite 35"

Niveau 1

Catégorie	Expansion	Fonction	Niveau syntaxique	Plan énonciatif
GN Possessif	Quantifiant	Circonstant	Primaire dans principale	Plan principal

Niveau 2

Macro-syntaxe	Position	Interprétation	Rôle actanciel
Noyau	Mediane	Immédiate	Autre

Niveau 3

Introduction du référent-1	Nom du référent-1	Nombre-1
Association	équipe de Flers	Groupe flou

Comme tout avait brûlé, la mère, les meubles et les photographies de la mère, pour Fabre et le fils Paul c'était tout de suite beaucoup d'ouvrage : toute cette cendre et ce deuil, déménager, courir se refaire dans les grandes surfaces. Fabre trouva trop vite quelque chose de moins vaste, deux pièces aux fonctions permutoables sous une cheminée de brique dont l'ombre donnait l'heure, et qui avaient ceci de bien d'être assez proches du quai de Valmy.

Le soir après le dîner, Fabre parlait à Paul de sa mère, sa mère à lui Paul, parfois dès le dîner. Comme on ne possédait plus de représentation de Sylvie Fabre, il s'épuisait à vouloir la décrire toujours plus exactement : au milieu de la cuisine requiert des hologrammes que dégonflait le minindre imprécision. Ça ne se rend pas, soupirait Fabre en posant une main sur sa tête, sur ses yeux, et le dérangement l'endormait. Souvent ce fut à Paul de déplier le canapé convertible, transformant les choses en chambre à coucher.

Le dimanche et certains jeudis, ils partaient sur le quai de Valmy vers la rue Marseille, la rue Dieu, ils allaient voir Sylvie Fabre. Elle les regardait de haut, tendait vers eux le flacon de parfum Piver, Forvil, elle souriait dans quinze mètres de robe bleue. Le grif d'un scoupirait tracé à la hanche. Il n'y avait pas d'autre image d'elle.

L'artiste Flers était représentée sur le flanc d'un immeuble, juste avant le coin de la rue. L'immeuble était plus maigre et plus solide, mieux tenu que les vieilles constructions qui se collaient en grincant contre lui, terrifiées par le plan d'occupation des-sols. En manque de marquise, son porche saturé de moulures portait le nom. (Wagner) de l'architecte-sculpteur gravé dans un cartouche en haut à droite. Et le mur sur lequel, avec toute son équipe, l'artiste Flers avait peint pour figurer Sylvie Fabre en pied, surplombait un petit espace vert rudimentaire, sorte de square sans accessoires qui ne consistait qu'à former le coin de la rue.

The screenshot shows the GATE (General Architecture for Text Engineering) interface. On the left is a tree view of resources:

- GATE
- Applications (ANNE)
- Language Resources (GateTestRun.td_0002B, Testcorpus)
- Processing Resources
 - ANNE NE Transducer
 - ANNE OrthoMatcher
 - ANNE POS Tagger
 - Direction
 - Amount
 - Genie
 - Constrain
 - Instrument

The main window displays a list of annotations under the "Text" tab. The annotations are color-coded and include:

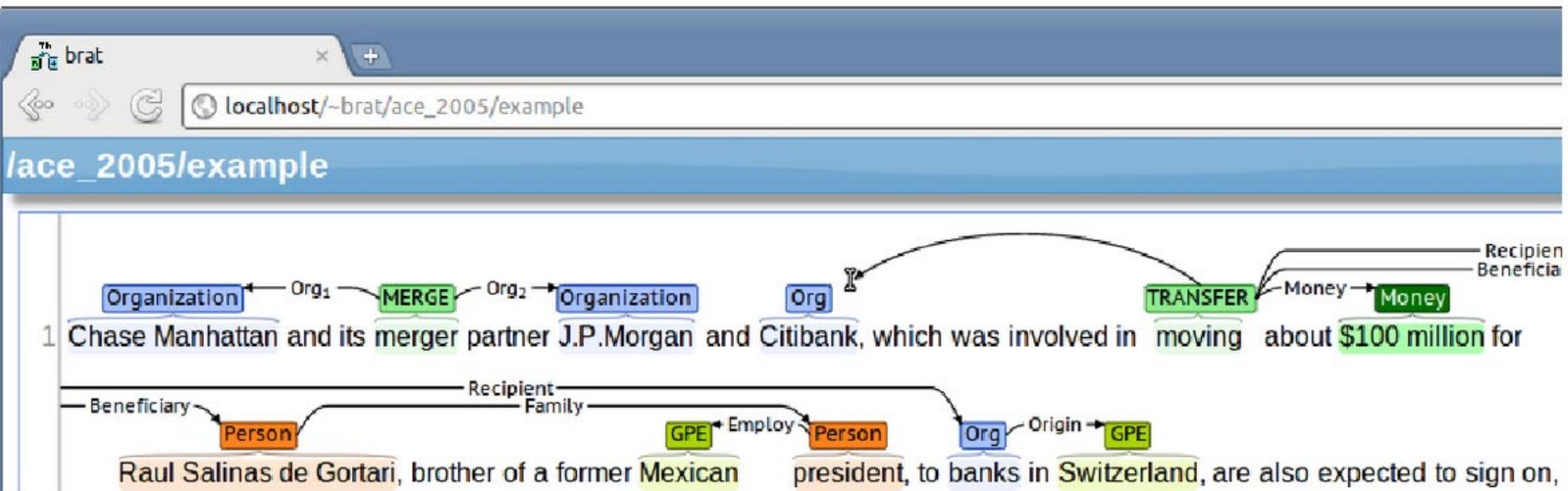
- Make it sound more **tappy** but don't make it too **thin**.
- Make the **tappy** timbre a touch **stronger**.
- Could you change the **airy** to **breathy** for the piano.
- Make the **drums** much more **bassy** but not to **boomy**.
- Change the **vocals** a bit from **delicate** to **edgy**.
- Make the **vocals** a bit more **edgy** and **match** less **thin**.
- Drum much **less** **body** and a little more **clear**.
- Make the **Drums** a touch more **tappy**.
- Make the **piano** a lot more **darker** without being **muted**.
- Make the **piano** **much** **less** **thinner**.
- Make the **acoustic guitar** **much** **less** **boomy** and a little more **delicate**.
- Make the **acoustic guitar** a bit more **brighter** but a lot less **harsh**.
- Make the **acoustic guitar** a bit more **warmer** but a bit less **woolly**.
- Make the **electric guitar** **much** **less** **harsh** and a bit brighter.
- Make the **electric guitar** a lot more **clear** and a bit more **airy**.
- Make the **electric guitar** **way** **less** **muffled** and a bit more **edges**.
- Make the **Snare** a bit more **woolly**.
- Reduce **vocal sibilance** a lot.
- Make **Vocal** a bit more **breathy**.
- Make the **vocal** some more **chesty** and a lot more **clear**.

On the right, a sidebar lists annotation types with checkboxes:

- Amount
- Constrain
- Direction
- Effect
- Instrument** (checkbox checked)
- Lookup
- Sentence
- SpaceToken
- SPLIT
- Timbre
- TimbreShift
- TOKEN
- UNKNOWN

At the bottom right, there is a section labeled "Original markups".

BRAT



BRAT

Although lymphokine genes are coordinately regulated upon antigen stimulation, they are regulated by the mechanisms common to all as well as those which are unique to each gene.

For most lymphokine genes, a combination of phorbol esters (phorbol 12-myristate 13 acetate, PMA) and calcium ionophores (A23187) is required for their maximal induction.

Yet phorbol ester alone or calcium ionophore alone produce several lymphokines.

The production of the granulocyte-macrophage colony stimulating factor (GM-CSF) is completely dependent on the two signals.

We have previously found a cis-acting region spanning the GM-CSF promoter region (positions -95 to +27) that confers inducibility to reporter genes in transient transfection assays.

Further analysis identified three elements required for efficient induction, referred to as GM2, GC-box and conserved lymphokine element (CLE0).

GM2 defines a binding site for protein(s) whose binding is inducible by PMA.

One protein, NF-GM2 is similar to the transcription factor NF- κ B.

GC-box is a binding site for constitutively bound proteins.

Choix d'outil

➤ Facilité d'installation et d'utilisation

- est-ce que le logiciel est disponible ?
- est-il facile d'installation, surtout pour des annotateurs peu familiers avec l'informatique ?
- l'interface utilisateur

Choix d'outil

➤ Besoins et contraintes de l'objet annoté

- si l'objet annoté nécessite d'annoter des relations, cette fonctionnalité doit être faisable à partir du logiciel
- format de sortie

➤ Fonctionnalités propres aux outils

- Accès aux ressources extérieures pour l'annotation
- Aspect collaboratif

Difficultés

➤ Outil

- Choix d'un outil adapté

- l'outil utilisé ne permet pas d'annoter les caractéristiques du phénomène (par exemple l'annotation de relation ou de discontinuités, de même que l'annotation avec des traits)
- détourner l'outil, cela signifie généralement complexifier l'annotation et amplifier les risques d'erreur.

- Difficulté de prise en main

Crowdsourcing ou myriadisation

➤ Les jeux ayant un but

- Phrase Detective

<https://anawiki.essex.ac.uk/phrasedetectives/>

- JeuxDeMots <http://www.jeuxdemots.org/jdm-accueil.php>

- ZombiLingoc <https://zombiludik.org/>

x10%

DONNER DES ASSOCIATIONS D'IDEES AVEC LE TERME QUI SUIT :

... record à battre de 300 Gr.

invité

Connexion pour plus de détails



carpaccio de viande

...

mettre un terme ici



Dernier terme proposé : [plat](#) • [supprimer](#)

Raffinements possibles :

1. [plat \(nourriture\)](#)
2. [plat \(pièce de vaisselle\)](#)
3. [plat \(horizontal\)](#)
4. [plat \(basculé\)](#)
5. [plat \(plongeant\)](#)
6. [plat \(uniforme\)](#)
7. [plat \(obsequieux\)](#)
8. [plat \(malchargé\)](#)
9. [plat \(balancé\)](#)
10. [plat \(minuscule\)](#)
11. [plat \(géométrique\)](#)
12. [plat \(nature\)](#)

Si vous ne savez pas répondre, il faut passer si partie. Si vous pensez qu'il n'y a pas de réponse possible, vous pouvez mettre """. Vous pouvez supprimer un mot proposé en cliquant dessus dans la liste affichée à droite.

5/10

[plat >>](#)
[camion >](#)
[manger >](#)
[nourriture >](#)
[viande >](#)

Interface du jeu JEUXDEMOTS.

niveau 4
50 310 / 100 000

283

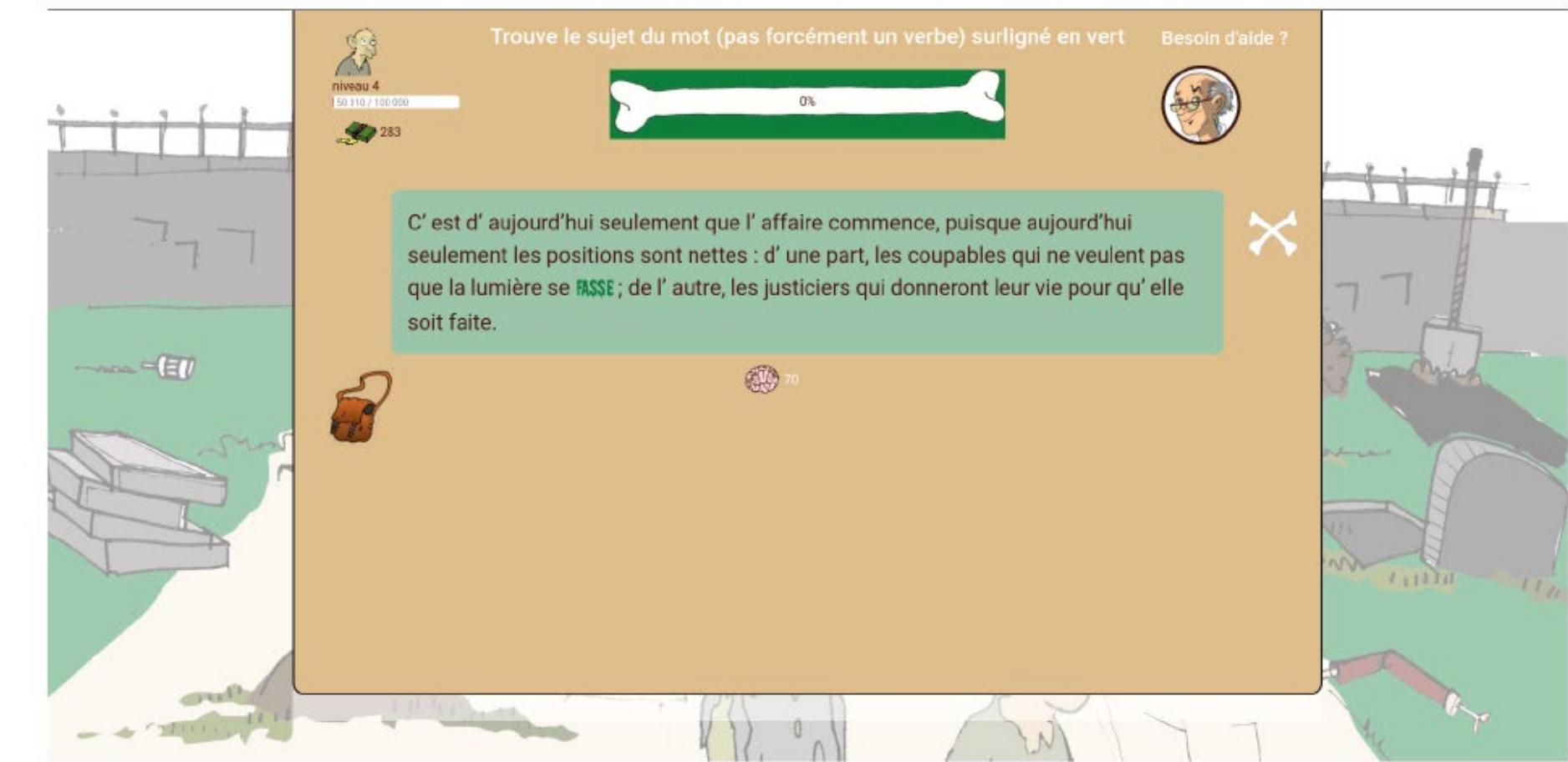
Trouve le sujet du mot (pas forcément un verbe) surligné en vert

Besoin d'aide ?



C' est d' aujourd'hui seulement que l' affaire commence, puisque aujourd'hui seulement les positions sont nettes : d' une part, les coupables qui ne veulent pas que la lumière se **FASSE**; de l' autre, les justiciers qui donneront leur vie pour qu' elle soit faite.

70



Exemple d'annotation sur le jeu ZOMBiLUDIK, jeu frère de ZombiLingo.

Types d'annotateurs

➤ les annotateurs **experts**

- **Expérience d'annotation**
- **Spécialiste d'un domaine ou d'un phénomène étudié**

➤ les annotateurs **non-experts**

Nombre d'annotateurs

- plus la tâche est complexe, plus il faudrait d'annotateurs

Guide d'annotation

Fort et al. (2009) recommandent :

- de définir précisément les termes, les catégories et justifier les choix effectués ;
- d'ajouter des exemples ;
- d'intégrer les potentielles ambiguïtés ;
- de préciser l'objectif de l'annotation ;
- de laisser une part d'interprétation pour les annotateurs

Le guide d'annotation ne sera pas parfait dès la première version, et il doit évoluer au fil de la campagne.

Méthodologie d'annotation

➤ Une phase d'entraînement

- consiste souvent, pour les annotateurs, à annoter une petite partie du corpus, via l'outil d'annotation,
- sert à se familiariser avec l'environnement d'annotation
- sert à se familiariser avec la documentation (guide d'annotation, manuel du logiciel, etc.).
- sert à améliorer le guide en évaluant l'accord inter-annotateur et en analysant les cas de désaccord.

Méthodologie d'annotation

- réaliser les phases de l'annotation simultanément :
 - annoter le phénomène dans son ensemble et en une seule fois
- décomposer en plusieurs tâches, ou phases.
 1. réaliser la première tâche d'annotation ;
 2. évaluer les annotations alors produites ;
 3. (étape optionnelle) construire un *gold standard* pour cette première tache ;
 4. à partir de ces annotations de référence, réaliser la deuxième tâche ;
 5. évaluer les annotations de cette deuxième tâche ;
 6. (étape optionnelle) établir un *gold standard* pour cette deuxième phase ;
 7. et ainsi de suite.

Difficultés

➤ Modélisation du phénomène

- La perception du phénomène implique déjà un choix.
- il est parfois ardu, voire impossible, de s'affranchir d'un modèle et d'atteindre une neutralité
- la modélisation du phénomène ne demeure pas toujours fixe, elle peut évoluer pendant la campagne, selon les contraintes techniques ou le retour des annotateurs.
- deux campagnes annotant le même phénomène.
 - le type de tâche à réaliser ne sera pas forcément le même
 - les schémas d'annotation ne sont pas forcément harmonisés
 - analyse de sentiment,
 - {Positif;Négatif},
 - {Positif;Neutre;Négatif},
 - une échelle de valeur intégrant la valence ($\{-2;-1; 0; 1; 2\}$).

Difficultés

- Corpus
 - le texte brut ou scanné ne permet pas de rendre compte des mises en pages
 - une perte d'information peut être préjudiciable à l'annotation
- Le fichier de transcription
 - l'absence d'un accès au document original (un fichier sonore)
- Le choix du corpus
 - le phénomène linguistique étudié n'est pas présenté du pdv quantitative et qualitative selon le type, le domaines ou le genre du texte.

Evaluation de l'annotation manuelle

Motivation

Pourquoi ne pas faire entièrement confiance à un annotateur ?

- il n'a peut être pas complètement compris la tâche
- il peut faire des erreurs d'inattention
- la tâche d'annotation n'est peut être pas claire (problèmes d'ambiguïté ou d'interprétation)

Mesures

- Accord inter-annotateurs
- Accord intra-annotateur
 - on compare les annotations d'un seul annotateur sur un même jeu de données à des périodes différentes (par exemple, au début et au milieu du processus).
 - cette mesure permettrait de vérifier la reproductibilité des annotations,
 - cette mesure permettrait de mettre en lumière des annotateurs dont les annotations manqueraient de consistance, c'est-à-dire de cohérence dans les annotations d'items voisins, par manque d'expérience ou d'implication.

Accord inter-annotateurs

L'annotation humaine étant un processus d'interprétation, sa validité ne peut être réellement évaluée. On se réfère alors le plus souvent au calcul d'un accord inter-annotateurs (AIA) qui permet :

- de mesurer la fiabilité et la qualité des annotations produites
- de vérifier la compréhension des consignes d'annotations par les annotateurs
- de fixer une borne supérieure aux performances que l'on peut attendre d'un système automatique

Lors du calcul de l'AIA, une des annotations est considérée comme "*référence*" et les autres sont alors appelées "*hypothèse*".

Accord inter-annotateurs

- L'accord est surtout utile pour vérifier la **fiabilité** de la tâche d'annotation
- Trois types de fiabilité :
 - la **stabilité** (*stability*) d'un annotateur, grâce à l'accord intra-annotateur, pour vérifier que sa manière d'annoter est constante ;
 - la **reproductibilité** (*replicability*), grâce à l'accord inter-annotateur, si les annotateurs annotent de la même façon en travaillant indépendamment des uns des autres ;
 - l'**exactitude** (*accuracy*) : en plus d'observer des sources de désaccord intra- et inter-annotateur, cette méthode permet de mesurer l'écart par rapport à une référence s'il en existe déjà une.

Mesure de l'accord inter-annotateurs

Ces mesures s'appliquent sur les résultats d'accords et de désaccords entre deux annotateurs. Plus exactement, cela consiste à prendre les annotations produites par l'un des annotateurs et d'évaluer les annotations d'un autre annotateur en fonction de celles-ci.

Parmi les mesures les plus utilisées :

- le Kappa de Cohen

Mesures (*Kappa de Cohen*)

- normalise l'accord observé en fonction d'un accord aléatoire, c'est-à-dire dû au hasard.
- compare des annotations observées avec une *référence aléatoire* portant sur l'ensemble des unités qui auraient pu être annotées
- calcule un rapport entre la **probabilité d'accord** P_o entre deux annotateurs et la **probabilité d'un accord aléatoire** P_e

Le calcul du *Kappa de Cohen* k s'effectue de la manière suivante :

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Remarque : Les valeurs de probabilité P_o et P_e dépendent du nombre de vrais négatifs (nombre d'annotations absentes de la référence et de l'évaluation) qui n'est pas toujours simple à estimer.

Le Kappa de Cohen

Probabilité d'accord (P_o) :

- 50 annotations au total

(oui = étiquette oui, non = étiquette non)

- accord sur 20 Oui et 15 Non

$$P_o = \frac{20+15}{50} = 0.7$$

Probabilité d'accord simultané (P_e) :

- A_1 a dit Oui 60% (30/50) et Non 40% de fois
- A_2 a dit Oui 50% (25/50) et Non 50% de fois

$$P_e = (0.5 \times 0.6) + (0.4 \times 0.5) = 0,5$$

Calcul du kappa :

$$k = \frac{0.7 - 0.5}{1 - 0.5} = 0.4$$

		A_1	A_1
		Oui	Non
A_2	Oui	20	5
A_2	Non	10	15

Matrice de confusion

Mesures pour plus que 2 annotateurs

- Multi- π de Fleiss (Fleiss, 1971)
- Multi- κ (Davies et Fleiss, 1982)

Mesures avec une pondération des catégories

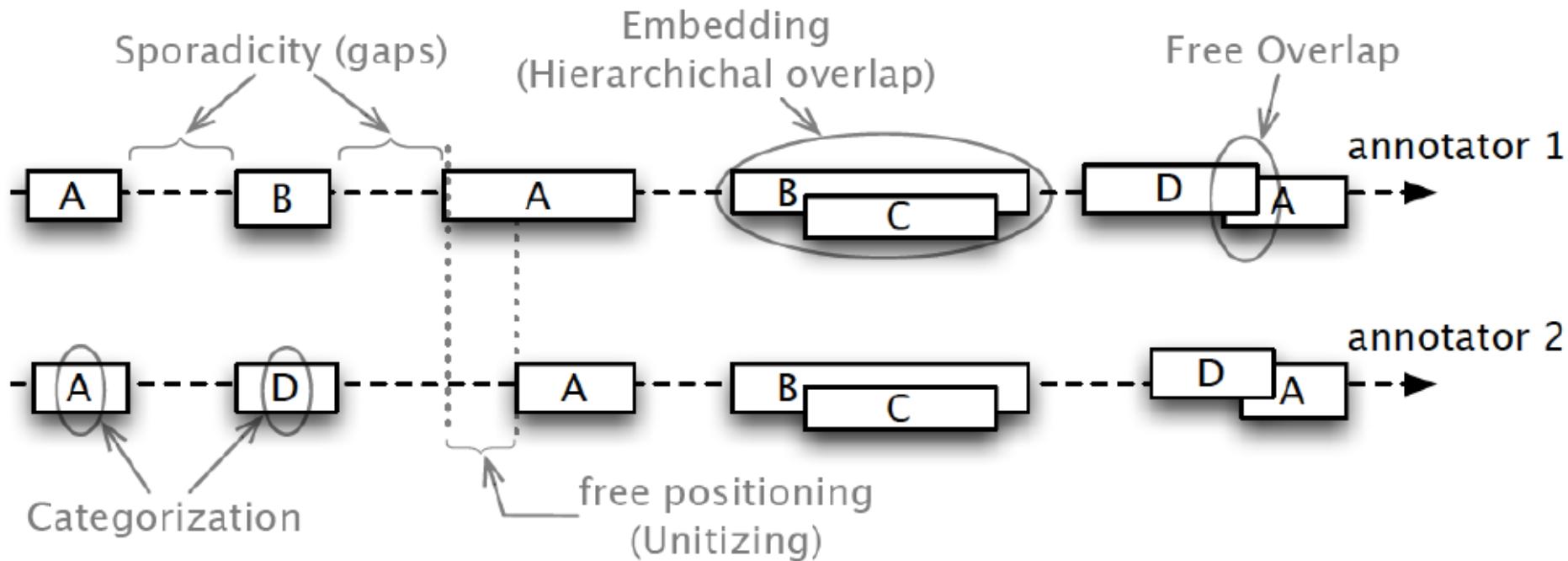
- tous les désaccords entre des catégories n'ont pas forcément la même importance :
 - en annotation morpho-syntaxique, assigner la catégorie « Nom propre » à un « Nom commun » peut être jugée comme une erreur moins grave qu'assigner la catégorie « Verbe » à cette même unité.
- Les mesures pondérées intègrent une notion de distance
 - Elles donnent un coût plus ou moins important aux désaccords entre les différentes catégories, en pénalisant moins sévèrement deux catégories proches, et inversement
 - Elles s'appuient sur le désaccord (et non l'accord, à l'inverse des précédentes mesures présentées)

α de Krippendorff Krippendorff (2013)

Mesures intégrant l'*unitizing*

➤ L'*unitizing*, (Krippendorff 1995)

- un type d'annotation qui regroupe deux tâches imbriquées : la délimitation des items à annoter et la catégorisation de ces derniers.



Mesures intégrant l'*unitizing*

➤ $u\alpha$ de Krippendorff Krippendorff (1995)

- l'accord est alors défini par la quantité de chevauchement entre des unités ayant une catégorie identique.
 - d'autres mesures proposées plus tard par l'auteur
- $|_u\alpha$: ce coefficient prend en compte seulement le chevauchement des unités, sans tenir compte des catégories qui leur sont assignées. Cette mesure est notamment utile pour évaluer uniquement l'alignement des unités entre les différents annotateurs, pour savoir s'ils sont d'accord sur le positionnement des unités ;
- $_{cu}\alpha$: cette mesure permet de mesurer le degré d'accord entre les catégories assignées aux unités repérées, tout en intégrant une pondération entre les catégories (la même que α). Toutefois, le calcul ne peut se faire que sur des unités qui se recoupent entre les ensembles des annotateurs ;
- $(k)_u\alpha$: présenté dans KRIPPENDORFF et al. (2016), ce coefficient mesure la fiabilité de chaque catégorie.

Mesures intégrant l'*unitizing*

➤ γ de Mathet, Widlöcher et Métivier

- Les auteurs considèrent que l'alignement et la mesure doivent être envisagées en même temps et ne devraient pas être traitées séparément lorsqu'on nous évaluons les annotations manuelles.

Interprétation de l'AIA

➤ savoir interpréter les valeurs :

- à partir de quel seuil d'accord les annotations peuvent-elles être considérées comme fiables ?
- la mesure d'accord dépend-elle de la tâche ?

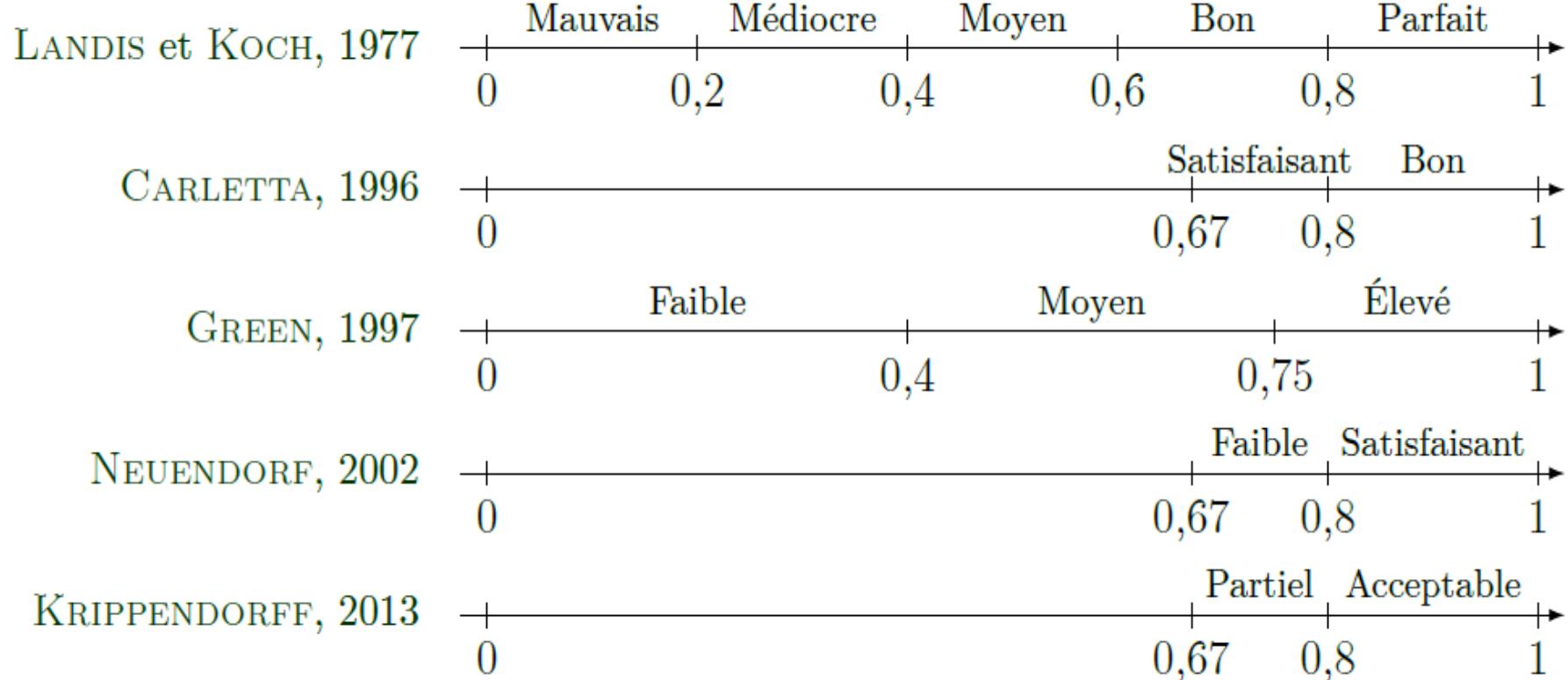
Interprétation de l'AIA

L'interprétation de l'AIA est elle aussi une tâche ardue.

Une des références a pour longtemps été l'échelle proposée par *Landis (1977)*. Cette échelle considère les valeurs d'AIA au dessus de 0.4 comme acceptables.



L'état de l'art actuel préconise de suivre une convention plus stricte qui considère les valeurs d'AIA supérieures à 0.8 comme signe d'accord et les valeurs en dessous comme signe de désaccord *Artstein (2008)*.



Seuils de fiabilité de l'accord inter-annotateur (pour la mesure κ).

Interprétation de l'AIA

Gut et Bayerl (2004) : ex de campagne d'annotation

- six annotateurs
- une tâche de transcription prosodique, en annotant selon plusieurs niveaux (segmentation en phrases, en mots, en syllabes, en intervalles vocaliques, consonantiques et pauses, l'indication des tons et des hauteurs).
- le κ de Cohen : le calcul des AIA des paires d'annotateurs pour chaque type d'annotation.
- un accord proche de 1 pour l'annotation des mots et des intervalles et des accords inférieurs à 0,4 pour l'annotation des syllabes et des tons.
=> la difficulté — voire l'impossibilité — de produire une échelle d'interprétation de l'accord inter-annotateurs qui se voudrait universelle, tant l'accord dépend de la tâche.

Interprétation de l'AIA

- Juger la difficulté d'une tâche : moins le score d'accord est élevé, plus cette tâche sera considérée comme difficile.
 - Ex : le processus d'annotation du corpus CLISTER (Hiebel et al., 2022) :
 - un échantillon de phrases a été annoté sans guide.
 - L' α de Krippendorff a été calculé : un faible accord (0,239).
 - la tâche a alors été jugée difficile.

Biais d'annotation : Typologie générale

➤ Tâche d'annotation :

- chaque tâche a ses spécificités et la combinaison de ces dernières implique une complexité unique

Biais d'annotation : Typologie générale

➤ Corpus

- Passage d'un format à un autre :
 - L'annotation s'effectue parfois sur du texte brut, format qui ne permet pas de rendre compte des mises en pages => une perte d'information
 - la scannérisation et l'océrisation ne restituant pas l'intégrité et les particularités (physiques, notamment) de l'oeuvre.
- Absence de l'accès au document source
 - les documents sources ne sont pas nativement sous format électronique
 - fichier sonore
- Corpus non représentatif

Biais d'annotation : Typologie générale

➤ Outils

- Uoutil non adapté
 - l'outil utilisé ne permet pas d'annoter les caractéristiques du phénomène (relation ou discontinuités, traits)
 - la scannérisation et l'océrisation ne restituant pas l'intégrité et les particularités (physiques, notamment) de l'oeuvre.
- Difficulté de prise en main
 - les documents sources ne sont pas nativement sous format électronique
 - fichier sonore

Biais d'annotation : Typologie générale

➤ Guide et schéma d'annotation

- Quantité du jeu d'étiquettes
- Schéma non hiérarchisé :
 - des catégories proches sémantiquement
- Catégories proches sémantiquement
- Cas spéciaux
 - Prévoir une étiquette pour les cas problématiques n'entrant pas dans les catégories prédéfinies
 - Définir dans le schéma l'indication de degré de certitude
 - Possibilité d'assigner plusieurs catégories pour la même unité des catégories proches sémantiquement

Biais d'annotation : Typologie générale

➤ Qualité rédactionnelle du guide

- exhaustif : regrouper tous les cas
- bien écrit
- formuler si...alors
- etc.

Biais d'annotation : Typologie générale

➤ Annotateurs

- Niveau d'expertise de la tâche
- Niveau de connaissance du domaine (le discours de spécialité)
- Niveau de formation
- Profil socio d'annotateur
 - Style d'écriture
 - Niveau d'études
 - Motivation
 - Affectés ou pas par le contenu => partie pris

Biais d'annotation : Typologie générale

➤ Processus d'annotation

- Pré-annotations du phénomène
 - gagner du temps et minimiser les gestes à effectuer sur l'outil
 - mais peut influencer les choix des annotateurs
- Subdivision de la tâche

dans un premier temps, nous demandons à l'annotateur de segmenter les unités, sans chercher à les catégoriser

- permet de réduire la charge cognitive liée à l'annotation
- mais segmentation ne sera pas forcément la même si l'annotateur devait les catégoriser dans la foulée

Biais d'annotation : Typologie générale

➤ Processus d'annotation

- Ordre des items
 - L'annotateur peut être influencé dans ses annotations par l'ordre dans lequel lui sont présentés les items à annoter.
 - Une catégorie est systématiquement présente en début de phrase et une autre en fin de phrase
- Distribution de catégorie
 - Une catégorie très fréquente vs des catégories rares
- Retour arrière
 - le fait d'avoir accès ou non au travers de l'outil aux précédentes annotations

Biais d'annotation : Typologie générale

➤ Processus d'annotation

- Poids du contexte
 - Annoter des unités « indépendantes », qui ne requièrent pas d'être interprétées au sein d'un contexte plus large, se révèle une tâche avec une charge cognitive moindre par rapport à une annotation où la taille du contexte est importante
- Temps d'annotation
 - l'annotateur effectue une longue session d'annotation, il y a le risque qu'il fasse davantage d'erreurs à cause de la fatigue

Biais d'annotation : Typologie générale

➤ Evaluation d'annotation

- Mesures d'accord inter-annotateurs non adaptées
- Biais de l'annotateur (**paradoxe du κ**)
 - le κ favorise les distributions déséquilibrées entre les annotateurs, *a contrario* de π et α : quand les annotateurs vont être en désaccord avec la distribution des catégories, alors cela va faire croître l'accord ;
- Prévalence des catégories
 - si la distribution des catégories est déséquilibrée, l'accord obtenu sera focalisé presque exclusivement sur la catégorie rare.
 - En effet, les mesures corrigées par la chance sont sensibles à l'accord sur les catégories rares.

Etablir une référence

➤ Méthodes pour établir une référence

- **Vote à l'unanimité (consensus) :**
 - seuls les items pour lesquels les annotateurs sont en parfait accord sont gardés.
- **Vote à la majorité relative :**
 - Pour chaque item, l'annotation de référence est définie par la catégorie ayant eu le plus de vote. La majorité est variable selon le nombre de catégories et le nombre d'annotateurs, et il n'y a parfois pas de majorité.
- **Révision collégiale par adjudication :**
 - Les annotateurs et un référent (généralement un expert) se réunissent afin de discuter des annotations produites, et particulièrement des items dont les annotations sont fortement en désaccord.

Vote à l'unanimité : problèmes

- Le principe est problématique :
 - seuls des items faciles (c'est-à-dire les items pour lesquels il ne semble pas y avoir de difficultés pour l'annotation) sont gardés, tandis que ceux dont l'annotation est plus délicate sont supprimés du corpus de référence final, alors qu' il s'agit souvent des cas les plus intéressants d'un point de vue linguistique
- entraîner un système sur un corpus constitué avec une vote à l'unanimité ne permettrait pas au système de catégoriser, voire simplement de repérer, des occurrences du phénomène.
- Lorsque nous excluons les cas difficiles, les systèmes est alors surévalué par la mesure de validité.

Vote à majorité : problèmes

- la majorité n'est pas toujours atteinte ou une égalité peut survenir.
- une manière de procéder est alors d'accorder des «poids» différents à certaines annotations, selon le degré d'expertise ou de confiance des annotateurs.

Diffuser le corpus



Formats des annotations : Annotation insérée (*inline*) (Fort, 2012)

- Les annotations produites sont directement incluses dans les textes sources, permettant par la même occasion de modifier les textes bruts si les annotateurs y détectent une erreur.

Annotation directement insérée dans le texte, repris de ANNODIS

(PÉRY-WOODLEY et al., 2011)

```
<structure>
<context type="before">... en plus dépendantes de l'aide et de l'investissement étrangers. </context>
<CT NbCar="913" startofs="11600" para="1" list="0" heading="-1" id="geop_28CT_coder2_1280479728969" file="geop_28">
<firstCOREF>L'IDE (investissement direct étranger)</firstCOREF>
<tags><tag nature="Im_COREFdef" start="11600">L'IDE (investissement direct étranger)</tag><tag nature="Ia_COREFpro" start="11715">il
</tag><tag nature="Im_COREFdef" start="11830">l'IDE</tag><tag nature="Im_COREFdef" start="11886">l'IDE par habitant</tag><tag nature="Ia_COREFdem" start="12001">ce ratio</tag><tag nature="Im_COREFdef" start="12127">l'IDE</tag><tag nature="Ia_COREFdef_R" start="12151">l'IDE</tag><tag nature="Im_COREFind" start="12273">des IDE investis</tag><tag nature="Ia_COREFdem" start="12329">Ce chiffre</tag></tags>
<segment schema="CT_coder2_1280479728969" start="11600" end="12513"><fullVersion><cue type="Im_COREFdef">L'IDE (investissement direct étranger)</cue> est nécessaire pour mettre en valeur les ressources de la région. Pourtant, <cue type="Ia_COREFpro">il</cue> reste encore très faible. Parmi les pays en transition, l'Asie centrale est le parent pauvre du point de vue de <cue type="Im_COREFdef">l'IDE</cue>. La BERD a calculé, sur la période 1989-1999, que <cue type="Im_COREFdef">l'IDE par habitant</cue> avait été de 668 dollars pour les pays d'Europe centrale et orientale. Pour les pays de la CEI, <cue type="Ia_COREFdem">ce ratio</cue> était près de cinq fois inférieur, s'élevant à 140 dollars. Si on excepte le Kazakhstan qui a attiré près de 80 % de <cue type="Im_COREFdef">l'IDE</cue> en Asie centrale, <cue type="Ia_COREFdef_R">l'IDE</cue> est inférieur à 50 dollars par habitant. Malgré les hydrocarbures et les métaux, l'Asie centrale n'a reçu que 0,3 % <cue type="Im_COREFind">des IDE investis</cue> dans le monde sur la période 1998-2000 <cue type="Ia_COREFdem">Ce chiffre</cue> était nul dix ans plus tôt mais seuls les pays en développement du Pacifique sud ont attiré moins de capitaux que les pays d'Asie centrale sur cette période de trois années.<break type="paragraph"/></fullVersion>
<shortVersion>L'IDE (investissement direct étranger) est nécessaire ...</shortVersion>
</segment>

</CT>
<context type="after"> L'investissement est faible. Les pays de la région ont ainsi ...</context>
</structure>
```

Formats des annotations : Annotation déportée (stand-off) (Fort, 2012)

- Les annotations sont présentes dans un fichier séparé.

Annotation déportée. Exemple repris de ANNODIS

Texte brut :

Amélioration de la sécurité Le maire a invité les membres du conseil à élaborer le programme d'amélioration de la voirie communale et de la sécurité routière pour l'année 1999. Il a rappelé que plusieurs automobilistes ont quitté la chaussée à l'intersection de la RD192 et du chemin rural de la Vaux des Fossés et qu'il convient de modifier le régime de priorité à cet endroit. La pose d'un panneau stop paraît être la formule la mieux adaptée pour assurer la sécurité des usagers. En délibérant, l'assemblée a accepté la proposition du maire et l'a chargé de faire établir par les services de la DDE un dossier de demande de subvention dans le cadre de la répartition des amendes de police 1999.

```
<unit id="gold_1">
<metadata>
<author>gold</author>
<creation-date>1</creation-date>
</metadata>
<characterisation>
<type>UDE</type>
<featureSet>
<feature name="type">UDE</feature>
</featureSet>
</characterisation>
<positioning>
<start>
<singlePosition index="0"/>
</start>
<end>
<singlePosition index="27"/>
</end>
</positioning></unit>

<unit id="gold_30">
<metadata>
<author>gold</author>
<creation-date>30</creation-date>
</metadata>
<characterisation>
<type>UDE</type>
<featureSet>
<feature name="type">UDE</feature>
</featureSet>
</characterisation>
<positioning>
<start>
<singlePosition index="30"/>
</start>
<end>
<singlePosition index="70"/>
</end>
</positioning></unit>
```

Annotation déportée

Formats des annotations : une solution hybride

- le positionnement des unités est réalisé au contact du texte (*inline*) et la caractérisation est totalement ou partiellement séparée (*stand-off*).

Format hybride inline-stand-off.

SEQUOIA CANDITO et al., 2017.

Positionnement *inline*

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xmlns:txm="http://textometrie.org/1.0">
    <s>À peu près au même moment que <entite id="1">
        Gutenberg</entite> inventait l'imprimerie, <entite id="2">Gillet Bonnemire</entite> créait en 1450 la première forge à <entite id="3">Saint-Dizier</entite>, à l'actuel emplacement du CHS. </s>
</TEI>
```

Caractérisation *stand-off*

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xmlns:txm="http://textometrie.org/1.0">
    <annotations>
        <annotation id="1">EN.Personne</annotation>
        <annotation id="2">EN.Personne</annotation>
        <annotation id="3">EN.Lieu</annotation>
    </annotations>
</TEI>
```

Formats des annotations : linéaire

Les annotations sont incorporées au texte, séparées des unités par un symbole délimiteur: le Brown Corpus (Kucera & Francis, 1967), le Penn Tree

Scarlett_NAM joue_VER:pres avec_PRP le_DET:ART chat_NOM ._PUNCT

Formats des annotations : balisé

➤ Les annotations sont présentes dans des balises, encadrant les unités, et autorisant des structures hiérarchiques. Le format souvent utilisé est généralement une extension de XML.

```
<w      cat="NAM">Scarlett</w>    <w      cat="VER">joue</w>    <w  
cat="PRP">avec</w> <w cat="DET">le</w> <w cat="NOM">chat</w>  
<w cat="PUNCT">.</w>
```

Mettre à disposition

- plateformes dédiées
 - Corli, Ortolang, LRE Map, Clarin, ELRA ou encore LDC.
- le site du projet ou d'un chercheur qui y est associé
 - ESLO
- sur demande
 - FTB

Mettre à disposition

- le **corpus** et les **annotations** : ce sont bien sûr les fichiers les plus importants lors du processus de diffusion du corpus ;
- la **documentation** : il s'agit des documents tels que le guide et le schéma d'annotation, mais aussi tout document concernant les choix effectués durant le processus d'annotation et de l'établissement de la référence ;
- les **références relatives au projet** : indiquer *a minima* la référence à citer lorsque nous utilisons le corpus ; les références ayant aidé au projet peuvent aussi être utiles (outils, modèle linguistique adopté...).

Ex :

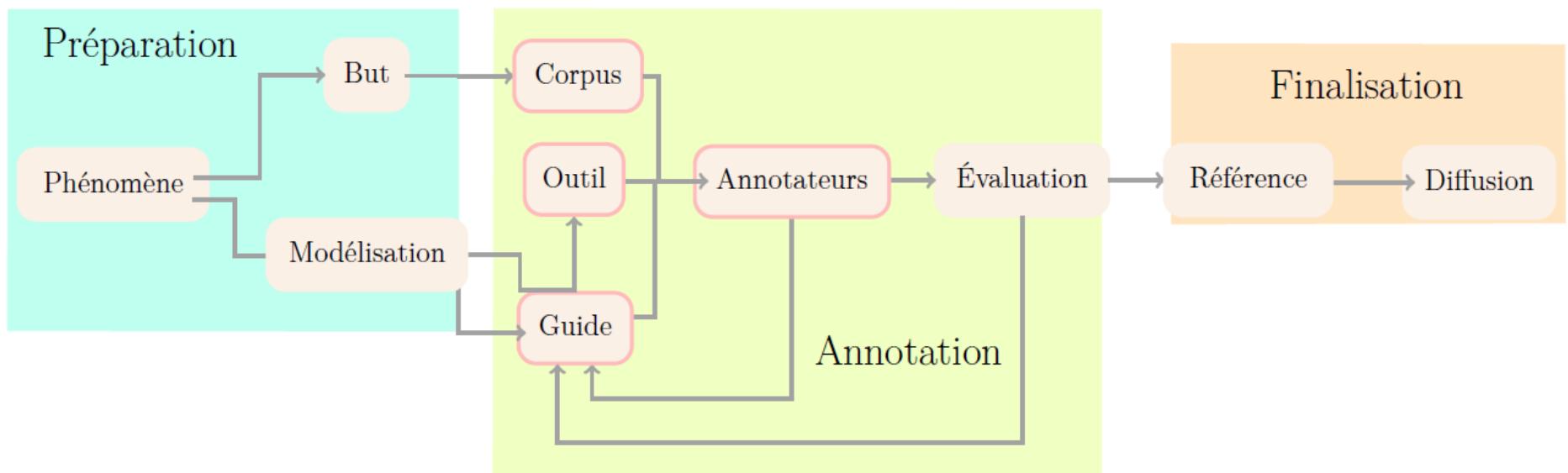
Le corpus ANCOR : <https://tln.lifat.univ-tours.fr/version-francaise/ressources/ancor-centre/corpus-ancor-centre-corpus-de-francais-parle-annote-en-coreference>

Mettre à disposition

- Données sensibles => anonymisation

Prix de l'annotation manuelle

- Prague Dependency Tree-Bank (Böhmová et al., 2003) => 600 000 \$ cinq ans du projet.
- le corpus Sequoia (Candito & Seddah, 2012)
=> 59 000 €.



Exemples

- **Unités déjà définies :**
 - les unités à annoter sont déjà délimitées et l'annotateur doit leur assigner une catégorie
- courriels pouvant être des spams (Metsis et al., 2006) ;
- textes relatifs aux transports ou non (Paroubek et al., 2018)
- phrases qui contient un segment obsolète (Laignelet, 2009) ;
- analyse de sentiment en positif/négatif
- Annotation en actes de dialogue : annoter les tours de parole selon leur fonction illocutoire
- POS
- désambiguïsation lexicale : assigner aux mots ambigus le sens selon le contexte
- Zoning argumentatif : assigner à chaque phrase une catégorie selon sa fonction argumentative

Phrase	Catégorie
Aujourd'hui, le PIB par habitant de la France est de 27 600 dollars.	Obsolète
En 2004, le PIB par habitant de la France est de 27 600 euros.	Non obsolète

Dialogue annoté selon le manuel d'annotation de ASHER et al.

(2017)

OPE	1	TC	Bonjour, je suis _TC1_, que puis-je pour vous ?
PRO	2	C	impossible pendant la lecture d'avancer la lecture
STA	3	C	_NUMTEL_
CLQ	4	TC	Si je comprends bien, le problème concerne la vidéo à la demande ?
STA	5	C	mais aussi l'enregistreur et la tv à la demande
INQ	6	C	pouvez vous m'appeler sur le portable ?
INQ	7	TC	Est ce que vous avez un message d'erreur ?
STA	8	C	non

Annotation prosodique, reprise de BUHmann et al. (2002)

he was there || and so was his girl-friend
I can tell you | this was un|be|lievable

|| représente les pauses fortes, tandis que | représente les pauses faibles

Segmentation thématique, extrait du corpus DEFT2006.

Le ministre de l'Agriculture, Christian Bonnet. C'est un élu breton, très actif et qui a été un excellent secrétaire d'état au logement, et qui va prendre en charge, à un moment difficile pour le fonctionnement du marché commun agricole, l'avenir de l'agriculture française. Le ministre du Travail, c'est M. Durafour, maire de Saint-Étienne, c'est-à-dire maire de la plus grande ville ouvrière de France, et qui a pu donc, dans la pratique de la vie municipale, connaître le monde du travail, sa représentation, ses problèmes, et qui établira, j'en suis sûr, les meilleurs échanges de vues possibles avec les travailleurs, leurs représentants et leurs organisations syndicales. Le ministre de l'Industrie est Michel d'Ornano, président du Conseil Régional de Basse-Normandie et qui a, je crois, les qualités d'organisation et d'efficacité nécessaires pour que nous poursuivions le développement de notre industrie, notamment le développement des créations d'emploi nécessaires pour assurer l'activité de la jeunesse française. J'ai tenu, avec le Premier ministre, à ce qu'il y ait, dans cette liste, pourtant restreinte, un ministre du Commerce et de l'Artisanat. Nous aurions pu l'appeler le ministre de l'Entreprise individuelle, nous avons gardé son titre traditionnel, et c'est M. Ansquer, vice-président du groupe UDR, à l'Assemblée nationale, et qui est aussi président du Conseil Régional des Pays de la Loire. À côté de cette liste, il y a deux nouveautés que je voulais vous signaler, deux autres ministres : d'abord une femme, Mme Simone Veil, qui est ministre de la Santé. Madame Simone Veil a été déportée avec sa famille à l'âge de 17 ans, à Ravensbruk ;

comprenant chacune une amorce et une clôture.

3. Fondements sociaux du concept en Occident

3.1 Les principes moraux

3.2 Le point de vue du droit

3.3 Le point de vue médical

3.4 Le point de vue psychologique

[...] C'est une notion assez vague, où l'on peut distinguer deux aspects :

- la maturité sociale, c'est-à-dire la capacité de [...]
- la maturité sexuelle, ou en d'autres termes la capacité

[...]

Ce qu'on peut en tout cas affirmer sur les deux alinéas précédents, c'est [...]

3.5 Rapprochements

Les approches explicités ci-dessus forment l'essentiel des principes qui justifient la manière dont nos sociétés perçoivent la pédophilie

SE1 Amorce

Item 1

Item 2

Item 3

Item 4

SE2 Amorce

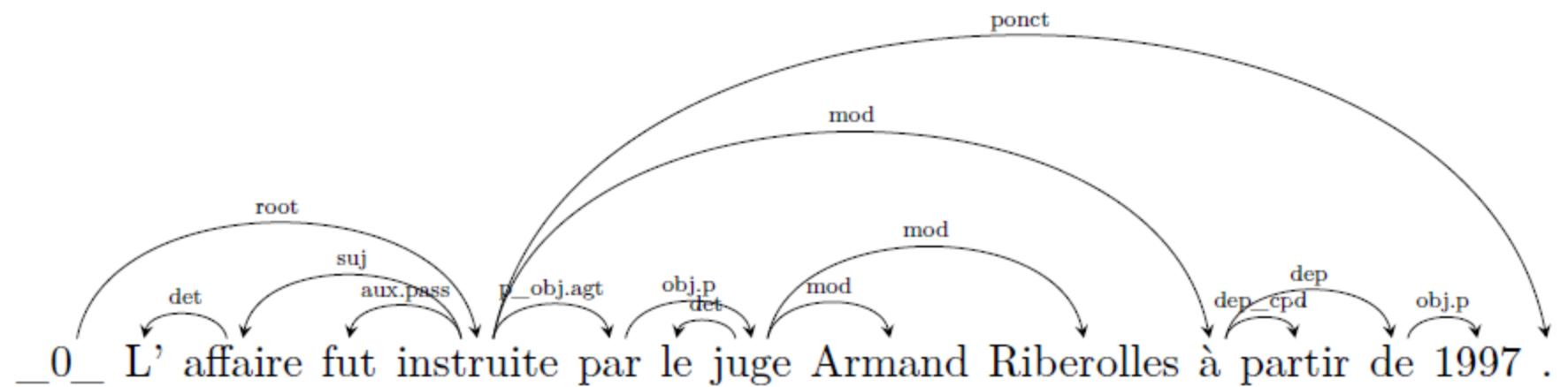
Item 1

Item 2

Clôture

Clôture

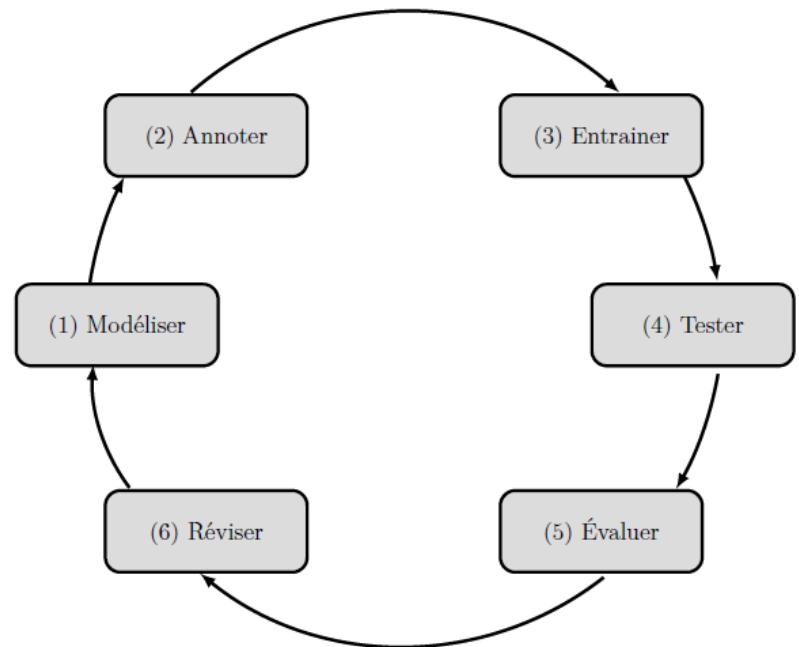
Annotation en syntaxe de dépendance



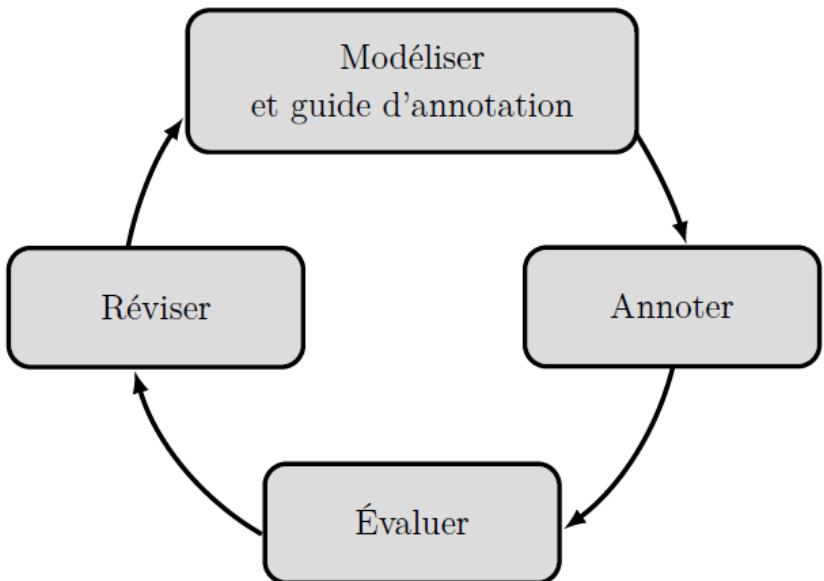
Campagnes d'évaluation

- Détermination du corpus et des formats
- Définition des objets étudiés
- Définition du modèle d'annotation
- Proposition de « feuilles de style » pour la visualisation des annotations
- Procédures de recherche et de vérification prédéfinies
- Rédaction du manuel d'annotation
- Analyse de l'accord inter-annotateurs
- Collecte et diffusion des données (corpus et annotations)

Annotation pour de l'apprentissage automatique



(a) Cycle MATTER



(b) Cycle MAMA

Cycles MATTER et MAMA repris de PUSTEJOVSKY et STUBBS (2012).