Fiche 6 - Exemple de protocole d'annotation

Sommaire

1. Présentation
2. Le protocole
   2.1. Segmentation
   2.2. Lemmatisation
   2.3. Etiquetage morpho-syntaxique
   2.4. Entités nommées
   2.5. Coréférence
   2.6. Attaques verbales

Mots-clés : annotation manuelle, protocole, couche d'annotation

1. Présentation

Une partie de ce projet a consisté à annoter manuellement des segments de textes sur INCEpTION. Dans l'objectif d'avoir les annotations les plus harmonieuses possibles et afin de faciliter la tâche d'annotation de chacun, nous avons décidé d'établir un exemple d'étapes d'une campagne d'annotation regroupant les lignes directrices. Le rôle de ce protocole est de servir d'appui pour les annotateurs, mais aussi de rendre compte des décisions prises face à certaines situations de doute pour chaque étapes.

Ce protocole a été défini avant le début des annotations afin de constater les capacités d'adaptation d'INCEpTION. Ensuite, au fil des annotations, des modifications ont dû être apportées sur nos lignes directrices car certaines n'étaient pas compatibles avec la plateforme. Dans cette fiche, nous présenterons uniquement les lignes directrices établies avant l'étape d'annotation, et nous renverrons ensuite vers la fiche dédiée à chaque couche pour détailler les problèmes rencontrés sur INCEpTION et les incompatibilités avec nos lignes directrices.

Nos lignes directrices et l'organisation des étapes d'annotation ont été réalisées à l'aide des conventions Universal Dependencies[1] et de la ressource GUM[2]. Notre protocole n'a pas pour vocation d'être une référence de l'annotation manuelle. Cependant, nous estimons important de l'inclure dans cette documentation afin d'aiguiller et d'inspirer de futurs annotateurs, tout comme nous avons pu être aiguillés et inspirés par la ressource GUM. Ce protocole s'organise selon les différentes couches d'annotations retenues pour ce projet.

[1] *https://universaldependencies.org*
[2] *Zeldes A. (2017). The GUM Corpus: Creating Multilayer Resources in the Classroom. Language Resources and Evaluation 51(3), 581–612.*

2. Le protocole

2.1 Segmentation :

Nos objectifs d'annotations prennent en compte une étape qui consiste à segmenter le texte en mot-forme (aussi appelés *tokens*). Cette étape permet d'isoler les unités lexicales, pour que par la suite, les annotations suivantes se basent sur la même unité.
Dans la plupart des cas, les unités lexicales sont séparées par des caractères d'espacement ou de ponctuation. Cependant certaines configurations peuvent être un peu moins évidentes à segmenter. Nous nous sommes donc accordés en amont sur certaines situations qui peuvent être sujettes à réflexion :

- Séparation des clitiques liés
  - sais-je → / sais / -je /
- Annulation des contractions du français
  - au → / à / le /
- Pas de segmentation pour les mots composés
  - arc-en-ciel → / arc-en-ciel
- Rattachement du tiret à la deuxième unité syntaxique
  - sais-je → / sais / -je /
- Rattachement des apostrophes avec la partie antérieure
  - l'avion → / l' / avion /

Pour plus de détails et pour l'application sur INCEpTION, voir Fiche n°9 - Lemmatisation.

2.2 Lemmatisation :

La lemmatisation consiste à attribuer pour chaque mot-forme la forme canonique qui lui est associée, c'est-à-dire une forme dénuée de toutes flexions. En cas de doute, la première forme de la catégorie morphologie du site du CNRTL. Le site d'Universal Dependencies propose également une liste non exhaustive des lemmes des mots grammaticaux.

Pour plus de détails et pour l'application sur INCEpTION, voir Fiche n°9 - Lemmatisation.

2.3 Etiquetage morpho-syntaxique :

L'étiquetage morpho-syntaxique consiste à attribuer une étiquette à un mot-forme qui correspond à sa classe grammaticale. Les mots-formes doivent être étiquetés selon le jeu d'étiquettes suivant :

Etiquette	Signification	Exemples
ADJ	*Adjectif*	*grand, petit*
ADP	*Préposition*	*après, avant, dans, en, chez*
ADP + DET	*Article contracté*	*du, au des, aux*
ADP + PRON	*Préposition + Pronom ou Pronom relatif*	*auquel, duquel, lequel*
ADV	*Adverbe*	*rapidement*
AUX	*Auxiliaire*	*être, avoir*
CCONJ	*Conjonction de coordination*	*mais, ou, et donc, or, ni ,car*
DET	*Déterminant*	*le, la, un, des, leur (s'il accompagne un nom)*
INTJ	*Interjection*	*hélas, chut, zut, oh*
NOUN	*Nom commun*	*chat, soleil*
NUM	*Numéral*	*1, deux, III*
PART	*Particule*	*-t, -ci, -là, ne*
PRON	*Pronom*	*leur (s'il remplace un nom)*
PROPN	*Nom propre*	*Selena Gomez, Nantes*
PUNCT	*Ponctuation*	*. ? , ! …*
SCONJ	*Conjonction de subordination*	*que, quand, comme, si, parce que, puisque, quand, qui, quoi*
SYM	*Symbole*	*#@*
VERB	*Verbe*	*manger, dormir*
X	*Mot étranger*	*sunshine*

Tableau 1 : Liste des étiquettes morpho-syntaxiques d'Universal Dependencies

Pour plus de détails et pour l'application sur INCEpTION, voir Fiche n°10 - Étiquetage morpho-syntaxique.

2.4 Entités nommées :

Les entités nommées sont des expressions linguistiques qui ont un référent unique et universel. Ce sont par exemple des expressions qui renvoient à une personne, à un lieu ou à une organisation.

D'après Universal Dependencies, les entités nommées doivent avoir été préalablement étiquetées comme "*nom propre*" (*PROPN*) ou selon les éléments qui la composent lors de l'étape de l'annotation morpho-syntaxique. Une même entités nommées peut être composée de plusieurs mots-formes.

Il faut être attentif à ne pas confondre les expressions référentielles et les entités nommées. L'entité nommée a un référent unique et stable, quels que soient le contexte d'interaction et/ou les interlocuteurs.

Trois principales étiquettes pour l'annotation des entités nommées :

- LOC → lieu
- ORG → organisation, marque
- PER → individu bien défini

Pour plus de détails et pour l'application sur INCEpTION, voir Fiche n°11 - Entités nommées.

2.5 Coréférence :

La coréférence est une annotation en chaîne. L'objectif est de relier tous les syntagmes nominaux qui ont le même référent. La coréférence sert à identifier le référent du discours mentionné par des expressions référentielles. Elle permet de suivre le comportement des référents du discours au fur et à mesure que ces expressions référentielles évoluent au cours d'un discours. La coréférence est importante puisqu'elle participe à la cohérence et la cohésion au sein d'un discours.

Afin d'annoter correctement la coréférence il est dans un premier temps essentiel d'identifier les expressions référentielles qui réfèrent à un même référent. Ainsi, plusieurs entités peuvent être marqueurs de coréférence comme :

- les noms propres
- les noms communs
- les pronoms

Pour les chaînes de coréférence, les coréférents sont rattachés à l'antécédent, et non pas à la première expression référentielle du référent.

Pour plus de détails et pour l'application sur INCEpTION, voir Fiche n°12 - Coréférence.

2.6 Attaques verbales :

La couche d'annotation des attaques verbales renvoie à l'identification et à la description des segments conflictuels. Ces segments peuvent être une phrase entière ou seulement une partie de cette phrase. Les attaques verbales forment une couche d'annotation spécifique à notre projet. Nous nous sommes basés sur les annotations proposées par Poudat & Ho-Dac (2019)[1].

Après l'identification des segments considérés comme étant des attaques verbales, il faut les décrire selon différents attributs :

- Le type de l'acte → c'est-à-dire ce à quoi le locuteur s'attaque
  - L'être : l'attaque porte sur ce que l'interlocuteur est
  - Le faire : l'attaque porte sur ce que l'interlocuteur a fait

- L'atténuation → c'est-à-dire, est-ce que les propos sont atténués ?
  - Description binaire → Oui si il y atténuation, sinon Non
  - Reconnaître les types d'atténuation possibles :
    - La concession : attaque nuancée par des connecteurs d'opposition (mais, au contraire, à l'opposé…).
    - La subjectivité exacerbée : l'auteur se pose comme la source évaluative de son énoncé en employant le "je" ou des marqueurs de point de vue "personnellement, pour moi, de mon point de vue". Celui-ci peut également feindre une incompréhension, des doutes ou exprimer sa surprise afin d'adoucir ses propos.
    - L'implémentation d'une suggestion qui peut se rapprocher d'un reproche.

- Le type d'acte illocutoire → quel est le sens réel du message au-delà du sens immédiat ?
  - Le reproche → "acte de langage évaluant négativement le comportement d'un individu et revendiquant son changement" (Kerbrat-Orecchioni, 1998)[2].
  - La menace → une forme d'avertissement ou de mise en garde, signe qui indique que quelque chose est à craindre
  - L'insulte → acte qui offense, blesse la dignité

- L'ironie
  - Description binaire → Oui si ironie, sinon Non
  - C'est le fait de faire entendre le contraire de ce qui est explicité

Pour plus de détails et pour l'application sur INCEpTION, voir Fiche n°13 - Attaques Verbales, et fiche n°8 - Création d'une couche d'annotation.

[1] Poudat, C., & Ho-Dac, L.-M. (2019). Désaccords et conflits dans le Wikipédia francophone. *Travaux linguistiques du CerLiCO*, 29.
[2] Kerbrat-Orecchioni, C. (1998). *L'implicite*. Armand Colin.