Plan séance du 3 décembre

1. Rappels sur le « pipeline »
2. Segmentation en phrases
3. Tokenisation

\+ évocation du fonctionnement de git

\+ éditeurs de texte

Sorbonne Nouvelle

Pipeline

texte brut $\rightarrow$ texte segmenté en phrases/paragraphes $\rightarrow$ texte tokenisé $\rightarrow$ texte lemmatisé $\rightarrow$ texte catégorisé $\rightarrow$ ...

- Annotation : ajout d'information (sans perte (?))
- Dépendance orientée entre les étapes
- Qualité : ces étapes préliminaires doivent être le plus près d'une annotation « *gold* » (référence)
- Mesure de la qualité : accord inter-annotateur

Sorbonne Nouvelle

Segmentation en phrases

entrée	texte brut (= suite de caractères)
sortie	texte découpé en unités (paragraphes/phrases)

Formatage : retour à la ligne (`\n`) ou séquence spécifique (`<s>`).

Sorbonne Nouvelle

Segmentation en phrases

entrée	texte brut (= suite de caractères)
sortie	texte découpé en unités (paragraphes/phrases)

Formatage : retour à la ligne (`\n`) ou séquence spécifique (`<s>`).

Algorithme : parcours du texte lettre par lettre, décision (contextuelle) : la lettre est un bon séparateur.

Sorbonne Nouvelle

Segmentation : manip

- Création d'un compte github (et demande par mail d'autorisation de participation)
- Choix d'un numéro d'extrait textuel inscrit dans le tableur partagé
- Application sur le texte choisi de la segmentation.
  Principe : une **phrase** par ligne et une ligne par phrase.
- Contribution au guide d'annotation : débat et synthèse
- Commit et commentaire

Sorbonne Nouvelle

Tokenisation

Token : terme volontairement neutre (vs. *mot*, *morphème*, *locution*...) pour désigner l'élément de base de l'annotation.

Les tokens sont séparés les uns des autres par des espaces, retour à la ligne, ou signes de ponctuation.

Problèmes :

- certaines coupes sont à éviter (“d'abord”)
- certains tokens sont à reconstituer (j' $\rightarrow$ je ; “du” $\rightarrow$ de+le)
- certains mots sont à regrouper (“tandis que”)

Sorbonne Nouvelle

Tokenisation : problèmes I

Apostrophe En français, l'apostrophe (a) sert à marquer une ellision → sépare 2 mots dont l'un est incomplet, et (b) apparaît de façon imprévisible dans quelques mots (1).

(1) d'abord, aujourd'hui, chef-d'œuvre, grand'mère, jusqu'à (?), main d'œuvre...

Statistiques : Emile ou l'éducation : 236 000 mots (28 000 formes) : 20 100 mots comprenant une apostrophe ;

24	aujourd'	914	c'	5706	l'	4706	qu'
3	chef-d'	1	ç'	15	lorsqu'	25	quelqu'
2	chefs-d'	3099	d' (dont abord)	397	m'	25	quoiqu'
10	grand'	1	donnez-m'	1	menez-l'	1576	s'
1	main-d'	698	j'	2642	n'	62	t'
39		149	jusqu'	28	puisqu'		

Sorbonne Nouvelle

Tokenisation : problèmes II

Tiret joue un rôle morphologique (forme un mot) ou syntaxique (forme une construction)

- **interne et imprévisible** : porte-monnaie (*vs* portefeuille), chou-fleur
- **interne et prévisible** : dix-sept, lui-même, avant-hier (?)
- **syntaxique** : voulez-vous, cet homme-là, c'est-à-dire
- **syntaxique + modifications morphologiques** : arrive-t-il, puissé-je

Sorbonne Nouvelle

Tokenisation : problèmes III

Locutions, expressions figées...

Le problème est de définir des critères. Gaston Gross 90, sur la seule construction N+Adj, recense jusqu'à 512 degrés de figement selon des critères linguistiques (distributionnels).

(2)
a. tandis n'existe pas tout seul $\rightarrow$ "tandis que"
b. je pensais alors que je devais agir (alors que)
c. encore que, au cours de, pomme de terre, chemin de fer
d. au moment (précis) où
e. Il a recouvert sa pomme de terre
f. Il est plus fort encore que je ne le pensais

Sorbonne Nouvelle

Tokenisation : problèmes IV

Ellisions

Le constituant ellidé peut être directement déduit dans certains cas, mais il y a des cas ambigus, et des cas où l'apostrophe ne marque pas d'ellision.

(3)
a. cas faciles : j', m', qu'
b. cas ambigus : d' (de/des), l' (le/la)
c. cas spéciaux : l'on aime (euphoniques)

Amalgames

cas contraire : au = à + le, du = (de + le —ou du) La question est souvent jusqu'où considérer qu'on a affaire à des amalgames (ex. quand = conj. temp + que)

(4) Quand tu viens et qu'il fait beau, je suis content

Sorbonne Nouvelle

Tokenisation : format

Aujourd'hui il boit du vin tandis qu'hier il buvait de l'eau.

```
# Aujourd'hui il boit du vin tandis qu'hier il buvait de l'eau.
Aujourd'hui
il
boit
de
le
vin
tandis que
hier
il
buvait
de
la
eau
.
```

Sorbonne Nouvelle