

Distant Reading 2: linguistique computationnelle

Une étape clef: préparation des données et entraînement

Simon Gabay



Lemmatiser

Définition

La lemmatisation désigne un **traitement lexical** apporté à un texte en vue de son analyse. Ce traitement consiste à appliquer aux **occurrences des lexèmes** sujets à flexion (en français, verbes, substantifs, adjectifs) un **codage** renvoyant à leur entrée lexicale commune (« **forme canonique** » enregistrée dans les dictionnaires de la langue, le plus couramment), que l'on désigne sous le terme de lemme.

Un exemple:

Texte	Les	étoiles	luisent	dans	la	nuit	noire	.
Lemme	Le	étoile	luire	dans	le	nuit	noir	.

Service en ligne

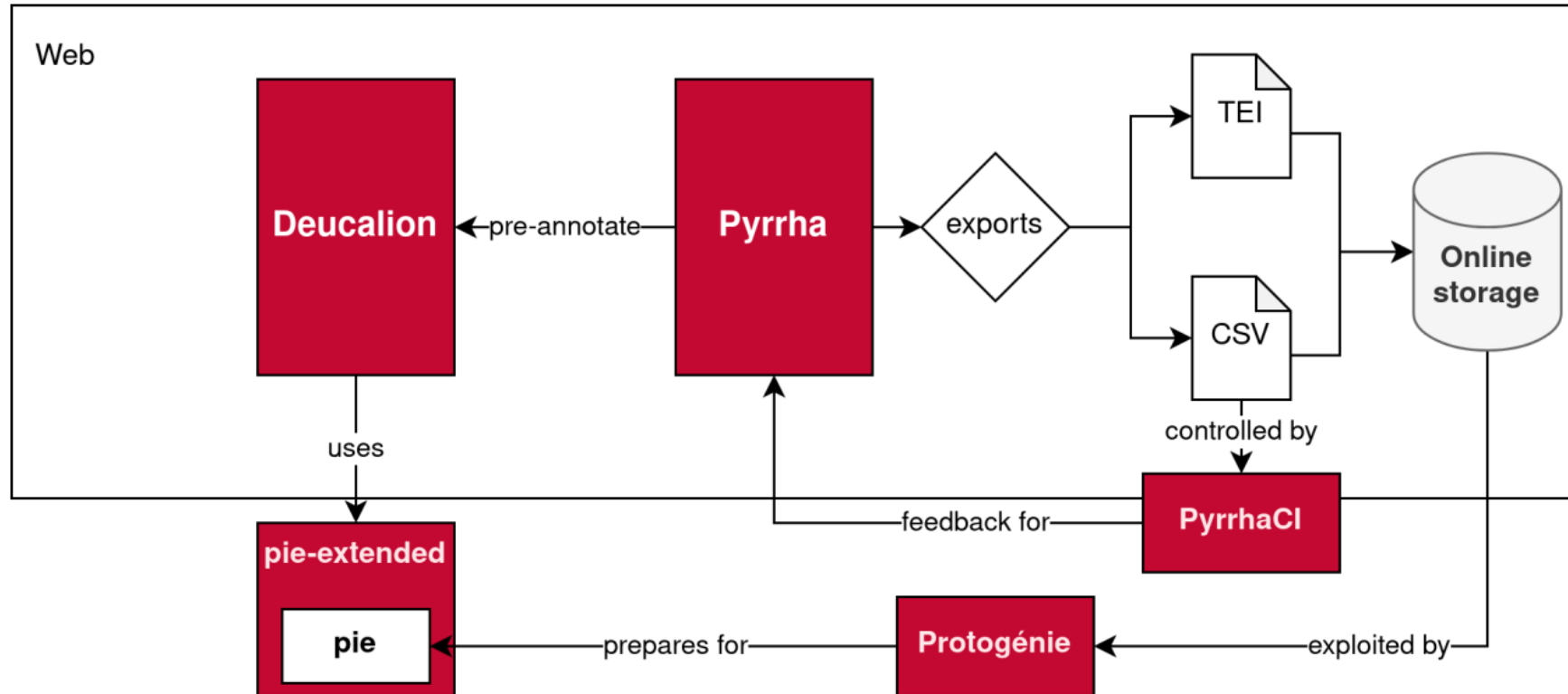
Il existe des services en ligne: <https://dh.chartes.psl.eu/deucalion/freem>

Vous pouvez faire un essai avec le texte suivant:

```
LA Cigale ayant chanté  
Tout l'Efté,  
Se trouva fort dépourvuë  
Quand la bife fut venuë.
```

Evidemment le résultat n'est pas magique. Il faut d'abord entraîner un modèle.

Cycle de travail



Thibault Clérice, Vincent Jolivet, Julien Pilla. "Building infrastructure for annotating medieval, classical and pre-orthographic languages: the Pyrrha ecosystem." *Digital Humanities 2022 (DH2022)*, Jul 2022, Tokyo, Japan. [hal-03606756](https://hal.archives-ouvertes.fr/hal-03606756).

Lemmatiser: le (pré-)problème de la tokenisation

- *pomme*
- *pomme de terre*
- *aujourd'hui*
- *c'est*
- *tire-bouchon*
- *veux-tu*
- *Celui-ci*
- *treshumble* (XVIIe s.)
- *C'est M. Dupont.*
- *bien que*
- *parce que*
- *ce pendant* (XVIe s.)

Lemmatiser: cas limites

Il y a un problème avec les homographes (impossibles à traiter avec une approche par lexique):

| *Il est à l'est de la ville.*

| *Je suis légaliste: je suis la loi.*

L'homographie peut être accidentelle (OCR, anciens états de langue):

| *Il va a Paris.*

Il peut aussi y avoir un problème avec le polymorphisme:

| *Il y a **besoing** et **besoin**.*

Lemmatiser: cas problématiques

- *comtesse*
- *va-t-il*
- *Jehan*
- *Jeanne*
- *Vespasianus*
- *François de La Rochefoucauld*
- *Oeuf*
- *Égypte*
- *aux*
- *dudit*

Lemmatiser: cas problématiques

- *comtesse* -> Féminin? ou masculin?
- *va-t-il* -> Que faire du -t- euphonique?
- *Jehan* -> normaliser les noms?
- *Jeanne* -> Féminiser les noms propres?
- *Vespasianus* -> Moderniser les noms?
- *François de La Rochefoucauld* -> *le* ou *La*?
- *Oeuf* -> comment traiter les ligatures?
- *Égypte* -> garder les accents sur les majuscules?
- *aux* -> lemme composé *à_le?* *à+le?*
- *dudit* -> lemme composé "triple" *de_le_dit?*

Lemmatiser: quelques cas

- *Il me demande à moi*
- *Un retour éclatant*
- *Une âme affligée*
- *Les .X. comandemenz (XVe s.)*
- *le pater noster*
- *s'enfuir*
- *Le Père R.*
- *il ne leur manquera rien vs leur vif éclat*
- *qui n'en veut? (corpus oral)*
- *Y z'y vont*

Lemmatiser: quelques cas

- *Il me demande à moi* -> moi=je ou moi?
- *Un retour éclatant* -> adj ou participe présent?
- *Une âme affligée* -> adj ou participe passé?
- *Les .X. commandements* -> que faire des chiffres?
- *le pater noster* -> comment lemmatiser les emprunts?
- *s'enfuir* -> que faire des pronominaux (*abaisser* vs *s'abaisser*?)
- *Le Père R.* -> Que faire des noms abrégés?
- *il ne leur manquera rien* -> pronom *il* vs déterminant possessif *leur*
- *qui n'en veut?* -> particule clitique de fausse liaison
- *Y z'y vont* -> transformation euphonique *il/s>y* sur le modèle de la liaison+ particule enclitique formant un pronom(?)

Entraîner un modèle

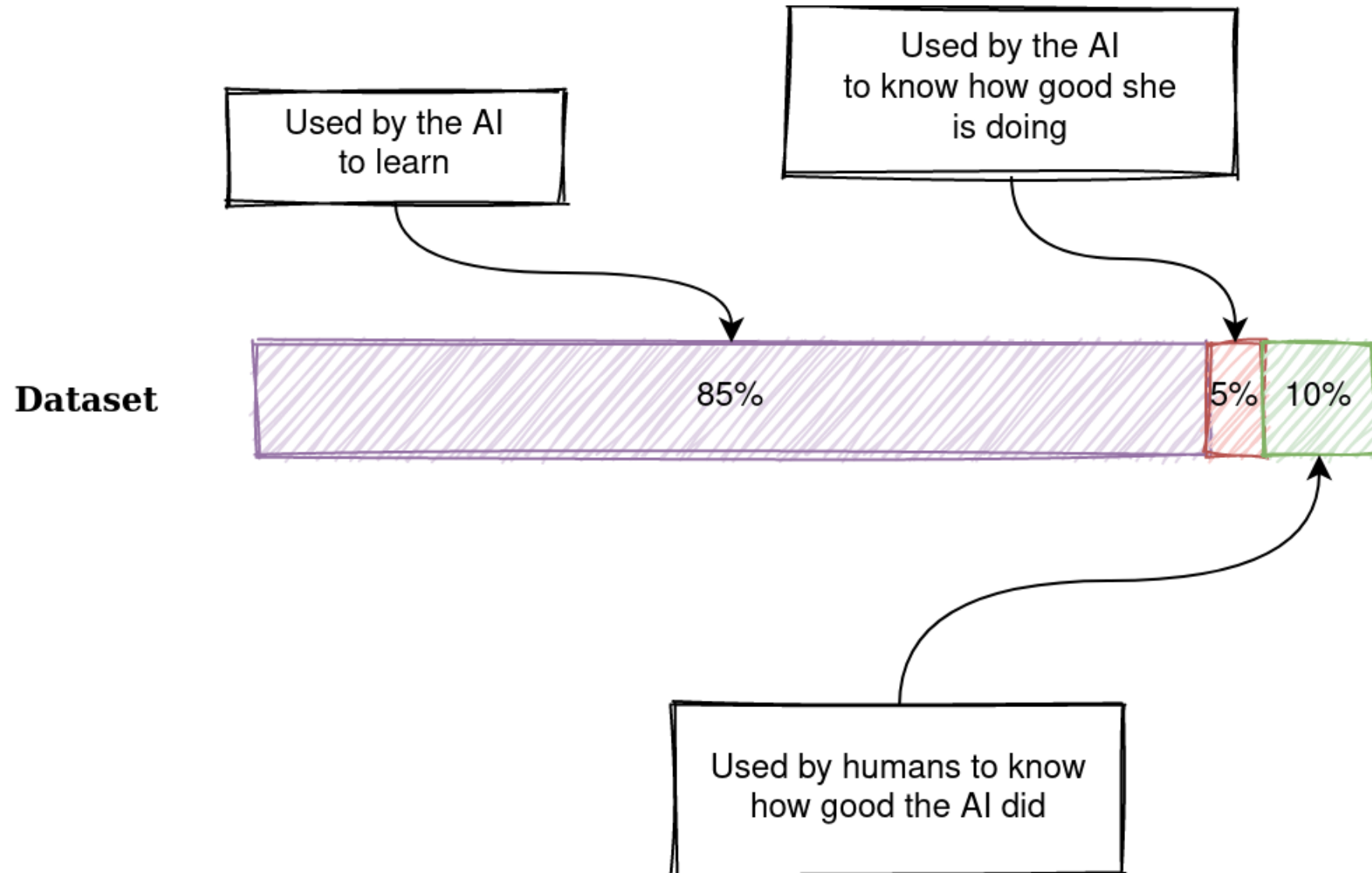
Annoter

Il va d'abord falloir préparer des données pour entraîner un modèle.

Il existe une interface pour accélérer la transcription (contrôle qualité, corrections en lot...):

<https://dh.chartes.psl.eu/pyrrha/>

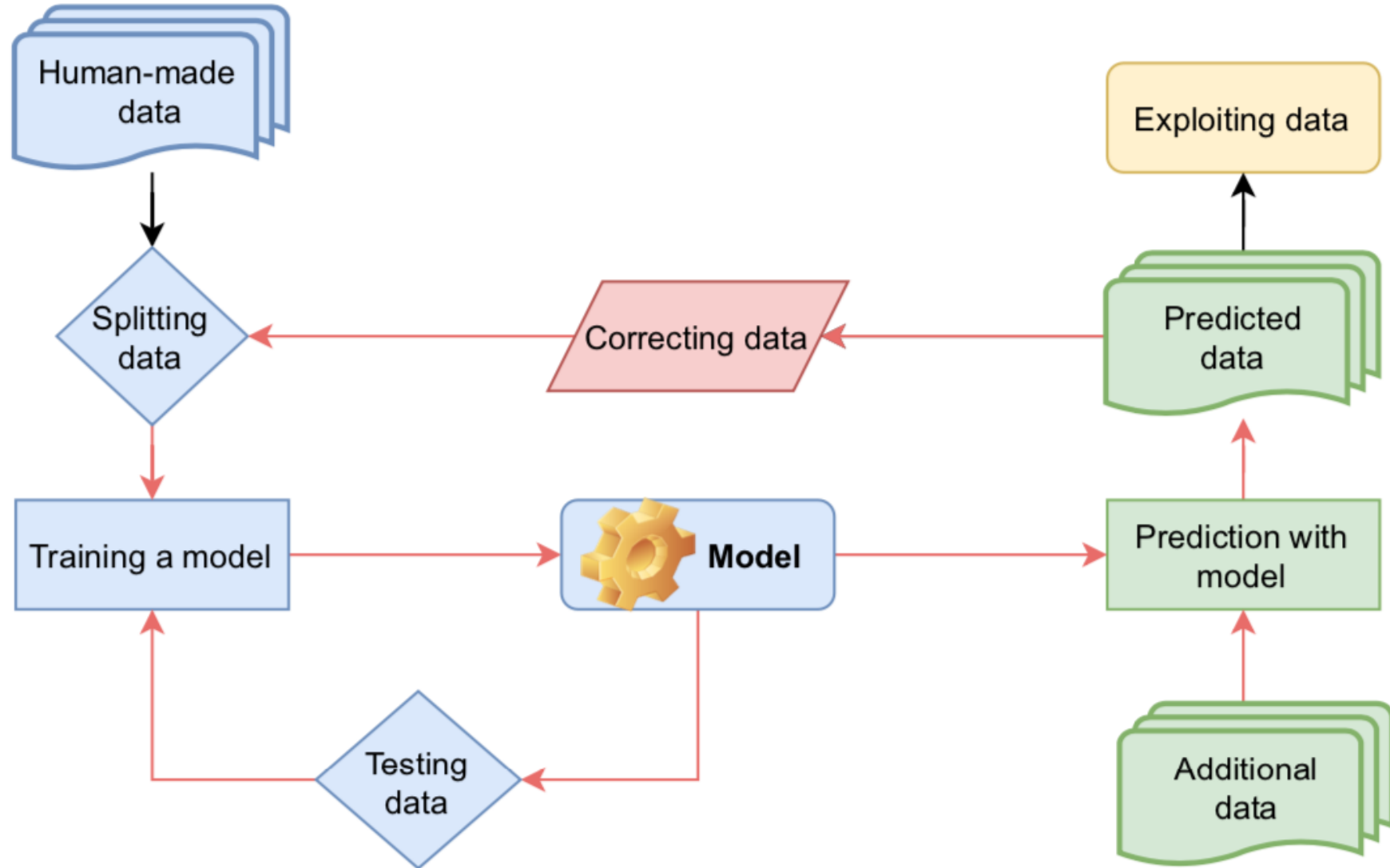
Faire le *split*



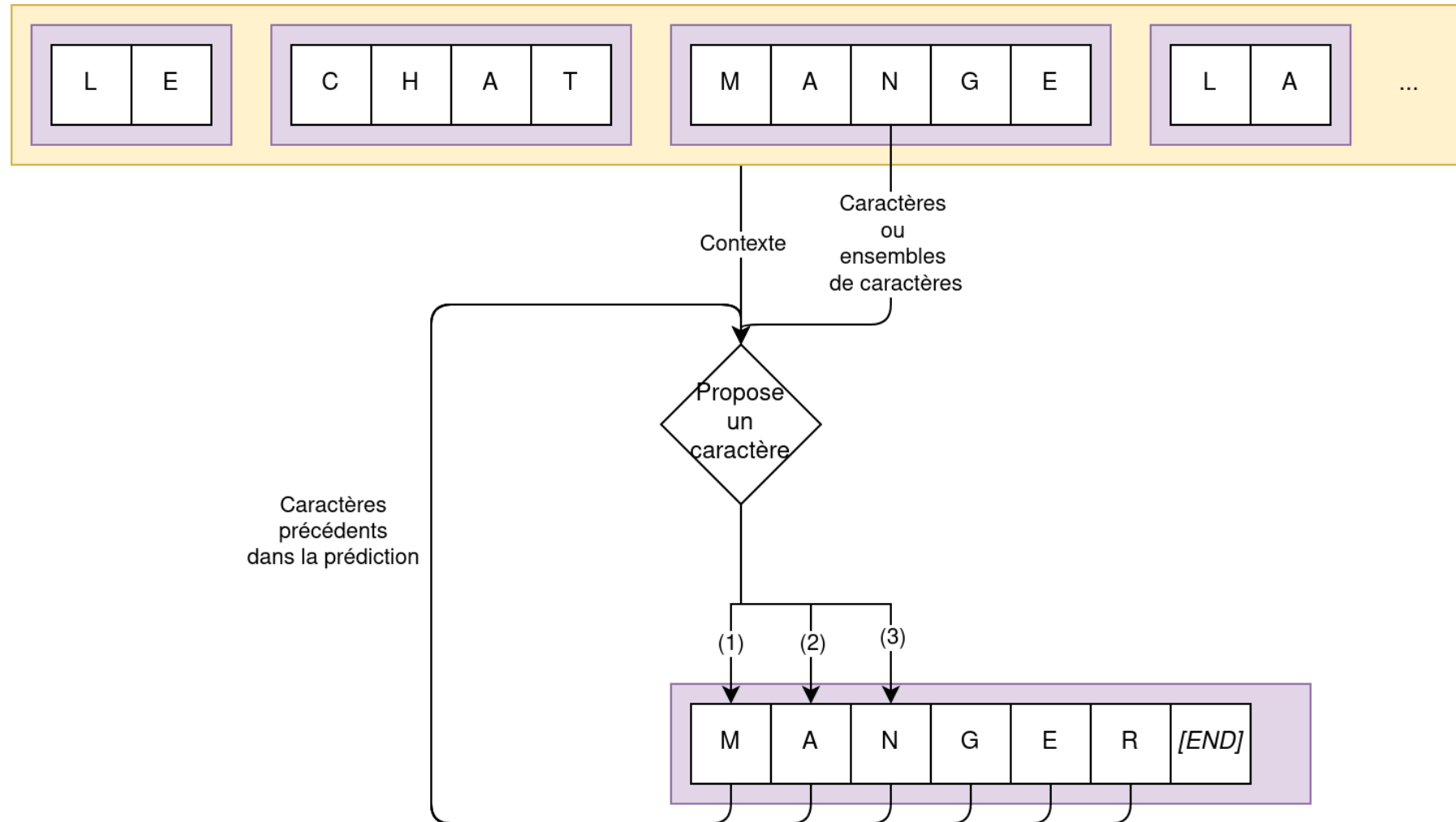
Faire le *split*: Protogénie

Il existe un outil pour faire ce split de manière propre: <https://github.com/hipster-philology/protogenie>

Cycle d'entraînement



Fonctionnement du lemmatiseur (génération de séquence)



Tester

Il faut impérativement tester le résultat. Pour cela on utilise des données

- *In-domain*, qui proviennent du même jeu de données que l'entraînement
- *Out-of-domain*, qui proviennent de données (plus ou moins) différentes de celles du jeu d'entraînement.

Parties du discours

Parties du discours: quelques cas

- *la Fortune*
- *un retour éclatant*
- *mon Dieu vs le dieu Jupiter*
- *parce que*
- *vive les vacances*
- *voici*
- *18 ans vs ses 18 ans*
- *le pater noster*
- *le Père R.*

Parties du discours: quelques cas

- *la Fortune* -> NOMpro ou NOMcom
- *un retour éclatant* -> VERppa
- *mon Dieu* (détermination non pertinente -> NOMpro) vs *le dieu Jupiter* (NOMcom)
- *parce que* -> ADVgen par analogie (*bien que*)
- *vive les vacances* -> VERcjjg
- *voici* -> VERcjjg
- *18 ans* DETcar NOMcom vs *ses 18 ans* DETpos ADJcar NOMcom
- *le pater noster* -> ETR
- *le Père R.* -> ABR

Parties du discours: quelques cas

- *aux*
- *duquel*
- *Monsieur de La Rochefoucauld*
- *Mesnil montant* (Ménilmontant)
- *là-dessus*
- *premier*
- *dernier*

Parties du discours: quelques cas

- *aux* -> PRE.DETdef
- *duquel* -> PRE.DETrel ou PRE.PR0rel ou PRE.PR0int en fonction du contexte
- *Monsieur de La Rochefoucauld* -> NOMcom PRE NOMpro NOMpro
- *Mesnil montant* (Ménilmontant) NOMpro VERppa
- *là-dessus* -> ADVgen PONfb1 ADVgen
- *premier* -> ADJord
- *dernier* -> ADJqua

Morphologie

Morphologie: quelques cas

- *je*
- *me*
- *moi*
- *mes*
- *vous êtes odieux*
- *il est clair*
- *J'ai mangé*
- *rien*
- *Julien Sorel*

Morphologie: quelques cas

- *je* -> CAS=n nominatif
- *me* -> CAS=r régime direct
- *moi* -> CAS=i régime indirect
- *mes* -> PERS.=1 | NOMB.=s ou PERS.=1 | NOMB.=p
- *vous êtes odieux* -> Féminin ou masculin? *vous* GENRE=x
- *il est clair* -> Masculin ou neutre? NOMB.=s | GENRE=n ou NOMB.=s | GENRE=m ?
- *J'ai mangé* -> VERcjg MODE=ind | TEMPS=pst + VERppe
- *rien* -> MORPH=empty
- *Julien Sorel*

Sources

- Thibault Clérice, Matthias Gille Levenson, Lucence Ing, Ariane Pinche, Simon Gabay, Jean-Baptiste Camps, «Lemmatiser des textes et corriger l'annotation grâce à l'apprentissage profond avec Pyrrha », *Humanistica 2021*, Mai 2021, Rennes, France. [⟨hal-03224112⟩](#).
- Simon Gabay, Jean-Baptiste Camps, Thibault Clérice. "Manuel d'annotation linguistique pour le français moderne (XVIe -XVIIIe siècles) : Version B." 2022. [⟨hal-02571190v2⟩](#).