Lyon, 26 septembre 2022
Séminaires IXXI − Traitement données complexes en Géographie

Ajout d'un étage d'analyse syntaxique à un corpus POS-tagué et lemmatisé de français classique

Benoît Crabbé − Pr U. Paris Cité / IUF
Achille Falaise − IE CNRS

Université de Paris

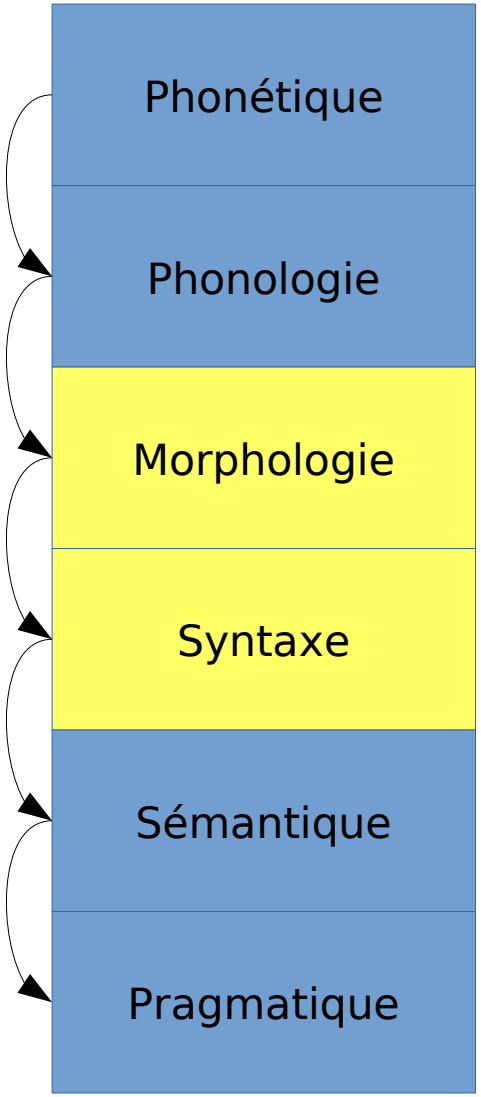
Laboratoire de linguistique formelle

cnrs

Plan

- Quelques termes et notions
- Projet Presto (2013-2017)
- HOPS & friends
- Projet Géode (2020-2024)

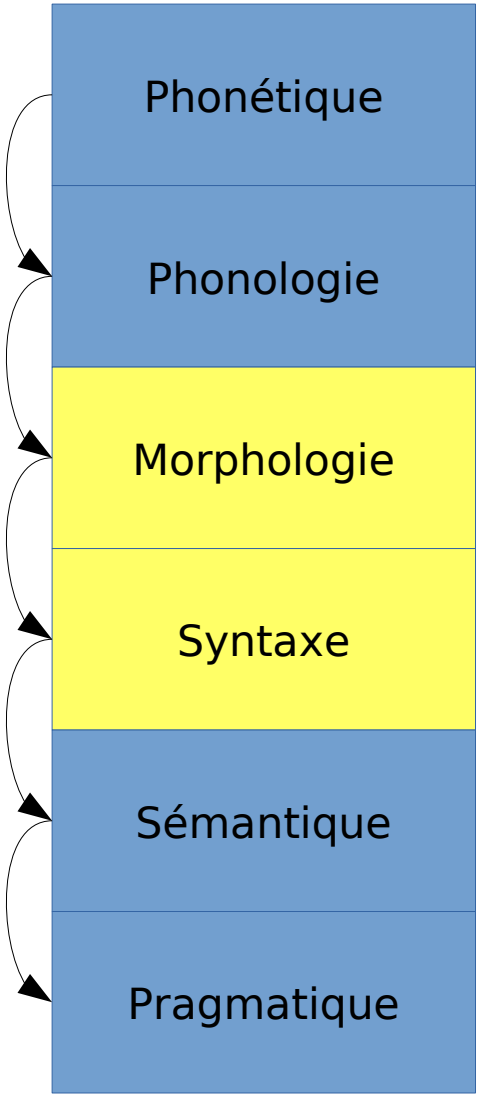
Quelques notions utiles
en Traitement Automatique des Langues (TAL)



Token ≃ Mot
Formes : *Disait* ; *lieux* ; *pour*
Lemme : *DIRE* ; *LIEU* ; *POUR*
Partie du discours (POS) : V ; N ; S
D'autres étiquettes...
→ *tagueur* / étiqueteur morphosyntaxique

Arbre
→ *parseur*

Quelques notions ut
en Traitement Automatiqu



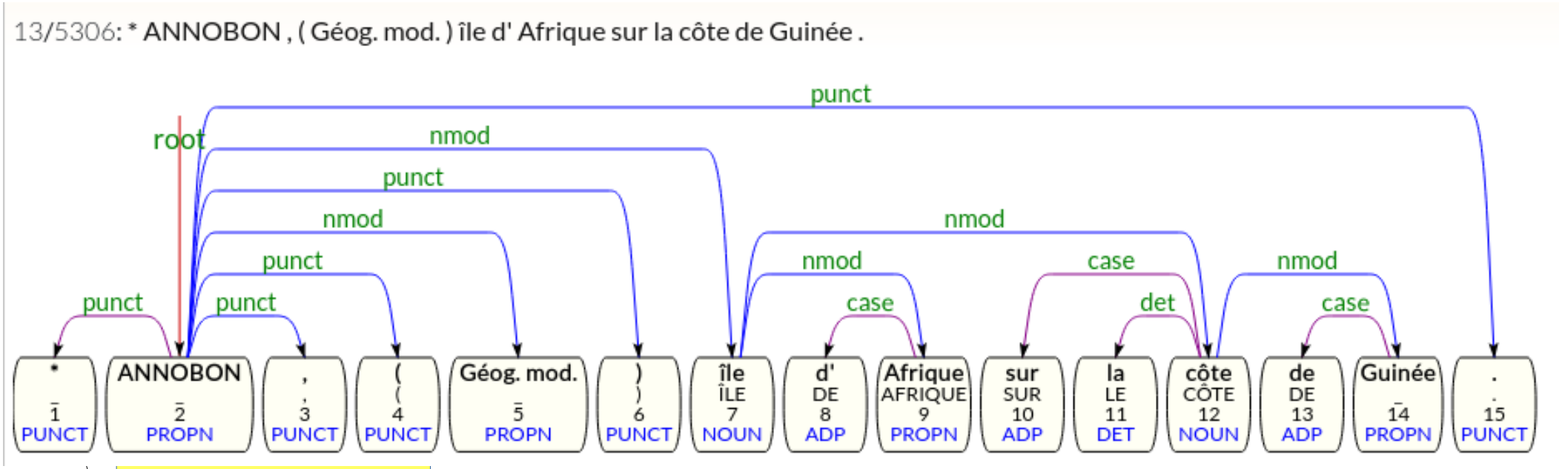
Token ≃
Formes
Lemme
Partie du
D'autres
→ *tague*

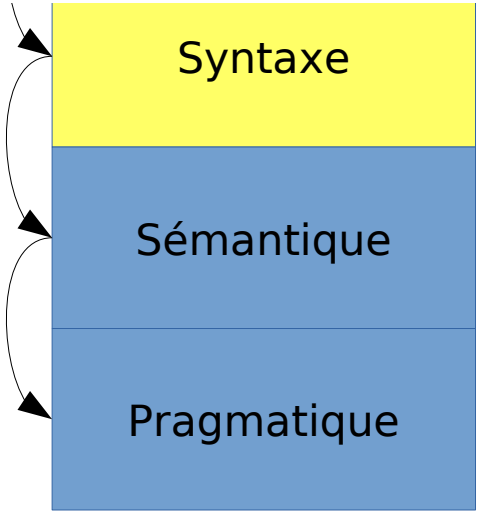
Arbre
→ *parse*

13/5306: * ANNOBON , (Géog. mod.) île d' Afrique

1	*	_	PUNCT
2	ANNOBON	_	PROPN
3	,	,	PUNCT
4	(	(	PUNCT
5	Géog. mod.	_	PROPN
6	)	)	PUNCT
7	île	ÎLE	NOUN
8	d'	DE	ADP
9	Afrique	AFRIQUE	PROPN
10	sur	SUR	ADP
11	la	LE	DET
12	côte	CÔTE	NOUN
13	de	DE	ADP
14	Guinée	_	PROPN
15	.	.	PUNCT

Quelques notions utiles
en Traitement Automatique des Langues (TAL)

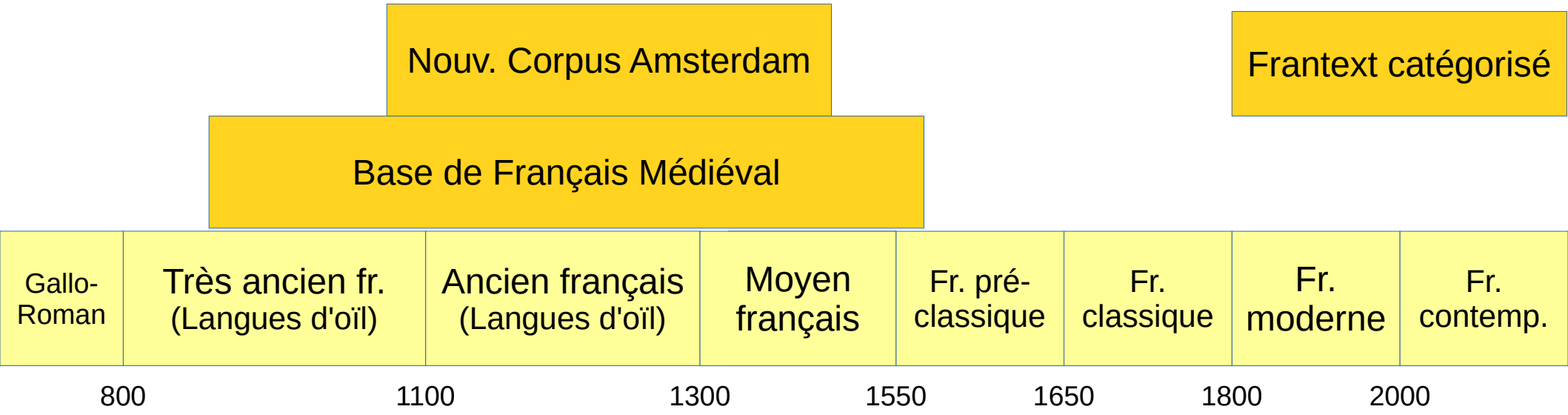




Arbre

→ *parseur*

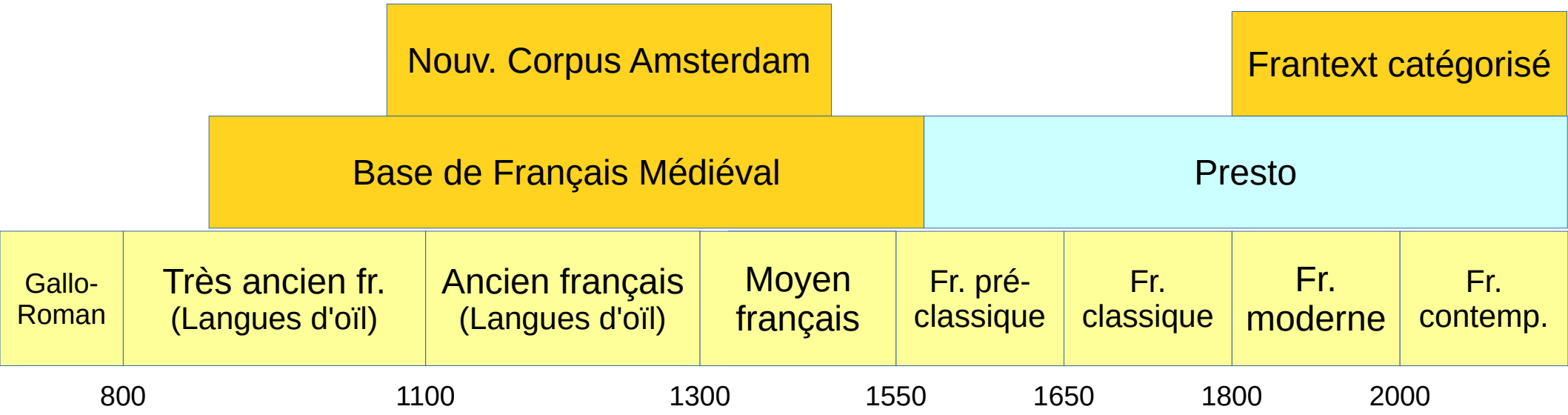
Projet Presto
2013-2017



Corpus historiques du français disponibles en 2013 avec au moins :

- ☑ Tokenisation
- ☑ Formes
- ☑ POS

Projet Presto
2013-2017



Corpus historiques du français disponibles en 2013 avec au moins :

- ☑ Tokenisation
- ☑ Formes
- ☑ POS

Le français du XVI^{e} siècle

SI LA NATURE (dont quelque Person- de grand'renommée non sans rayson a douté, si on la devoit appeller Mere, ou Maratre) eust donné aux Hommes un commun vouloir, & consentement, outre les innumerables commoditez, qui en feussent procedées, l'Inconstance humaine, n'eust eu besoing de se forger tant de manieres de parler. Laquéle diversité, & confusion, se peut à bon droict appeller la Tour de Babel.

La deffence, et illustration de la langue francoyse,
Joachim du Bellay, 1549.

Le français du XVIII^{e} siècle

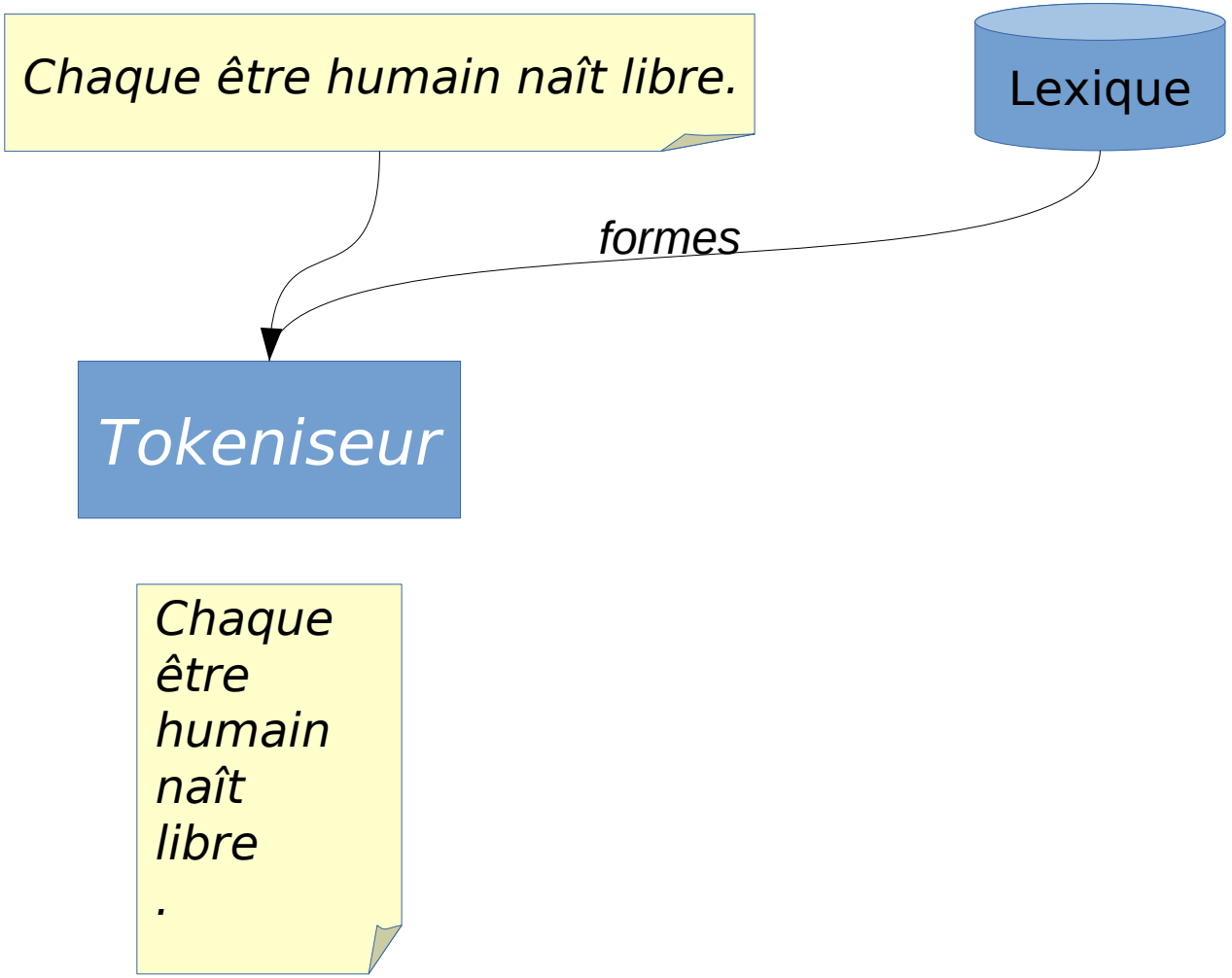
LANGAGE, s. m. (Arts. Raisonn. Philos. Metaphys.)

modus **&** usus loquendi, **maniere** dont les hommes se communiquent leurs pensées, par une suite de paroles, de gestes & d'expressions adaptées à leur génie, leurs **moeurs** & leurs climats.

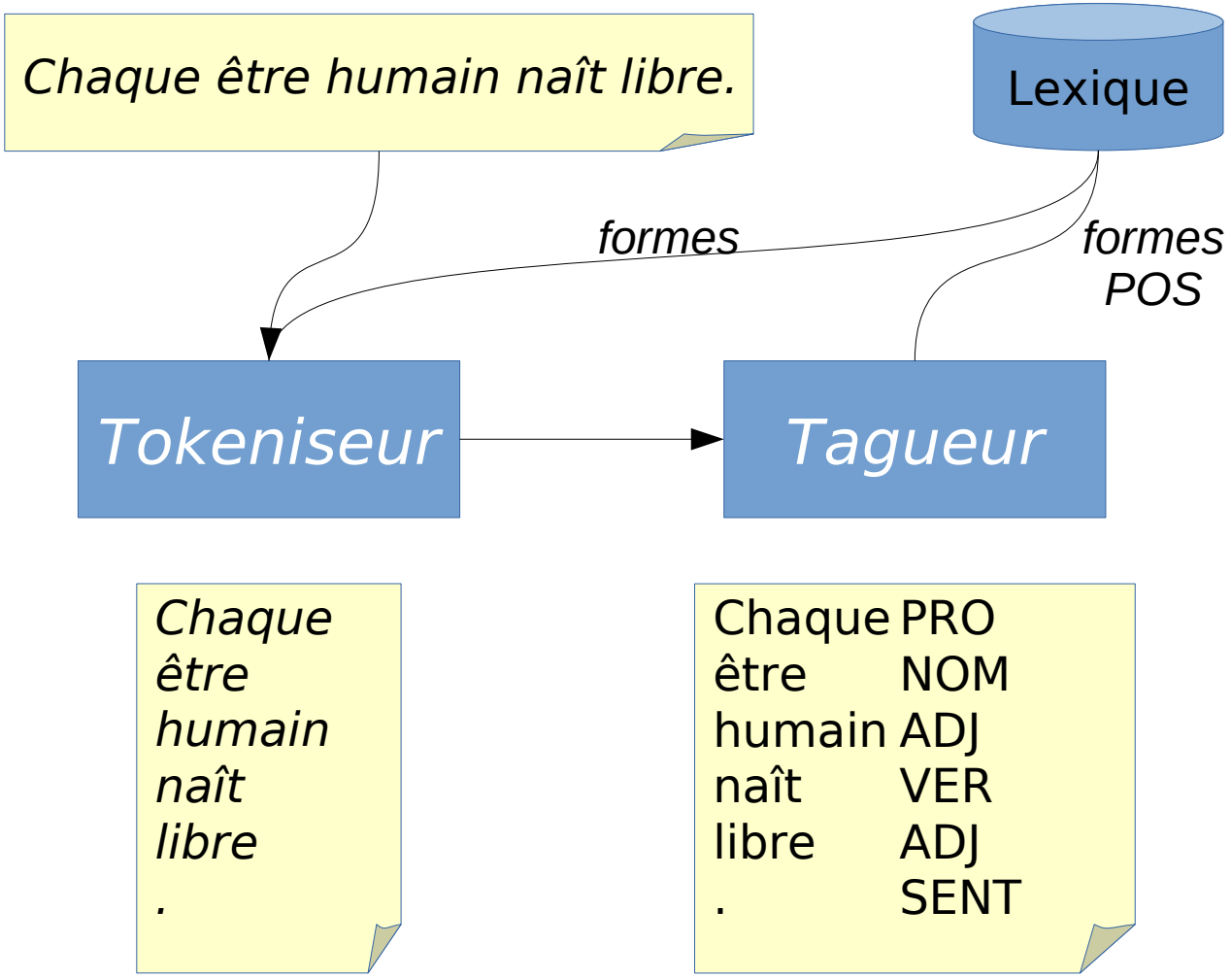
Dès que l'homme se sentit entraîné par goût, par besoin & par plaisir à l'union de ses semblables, il lui **étoit** nécessaire de développer son **ame** à un autre, & lui en communiquer les situations. Après avoir essayé plusieurs sortes d'expressions, il s'en tint à la plus naturelle, la plus utile & la plus étendue, celle de l'organe de la voix. Il **étoit aise** d'en faire usage en toute occasion, à chaque instant, & sans autre peine que celle de se donner des **mouvemens** de respiration, si doux à l'existence.

Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers,
Diderot & d'Alembert (dir.), 1751-1765.

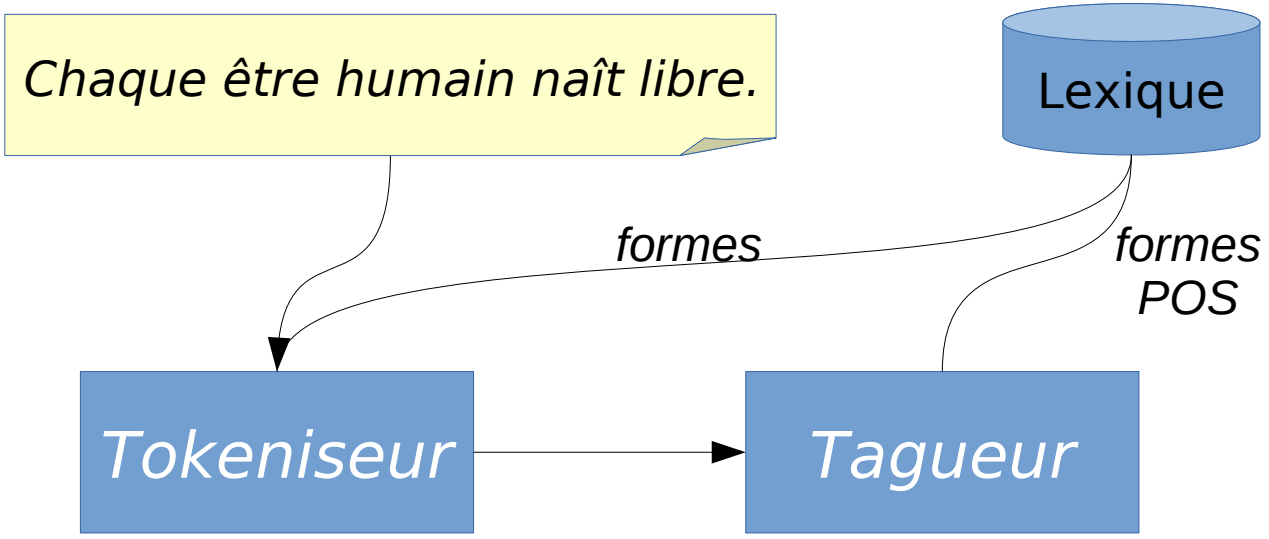
Chaîne de traitement Presto



Chaîne de traitement Presto

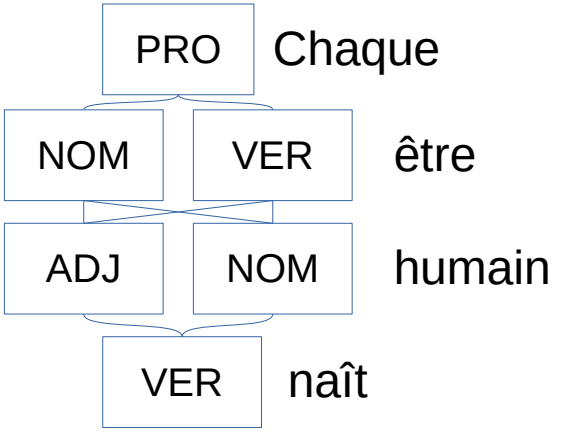


Chaîne de traitement Presto

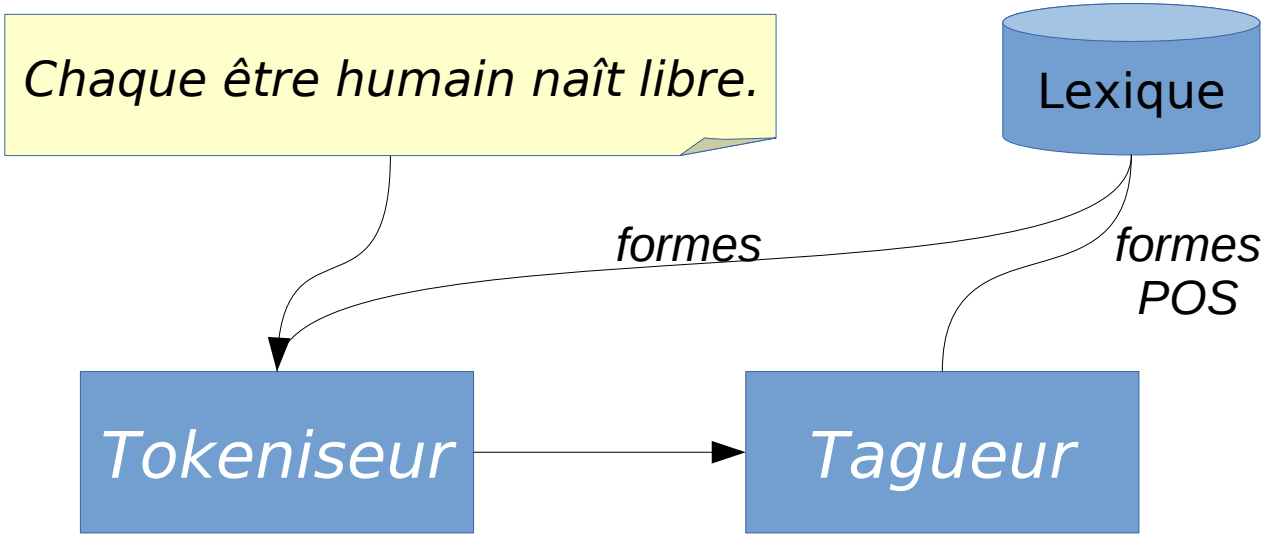


Chaque
être
humain
naît
libre
.

Chaque	PRO
être	NOM
humain	ADJ
naît	VER
libre	ADJ
.	SENT



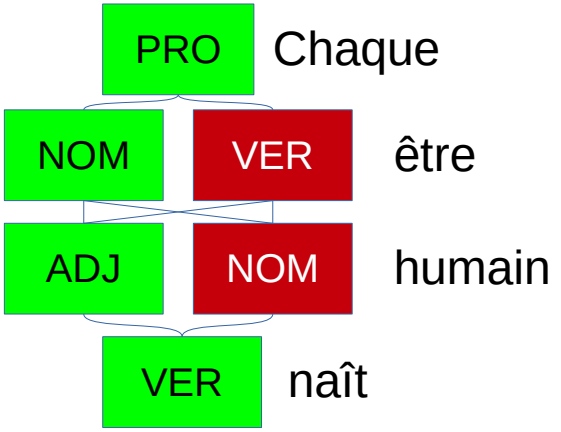
Chaîne de traitement Presto



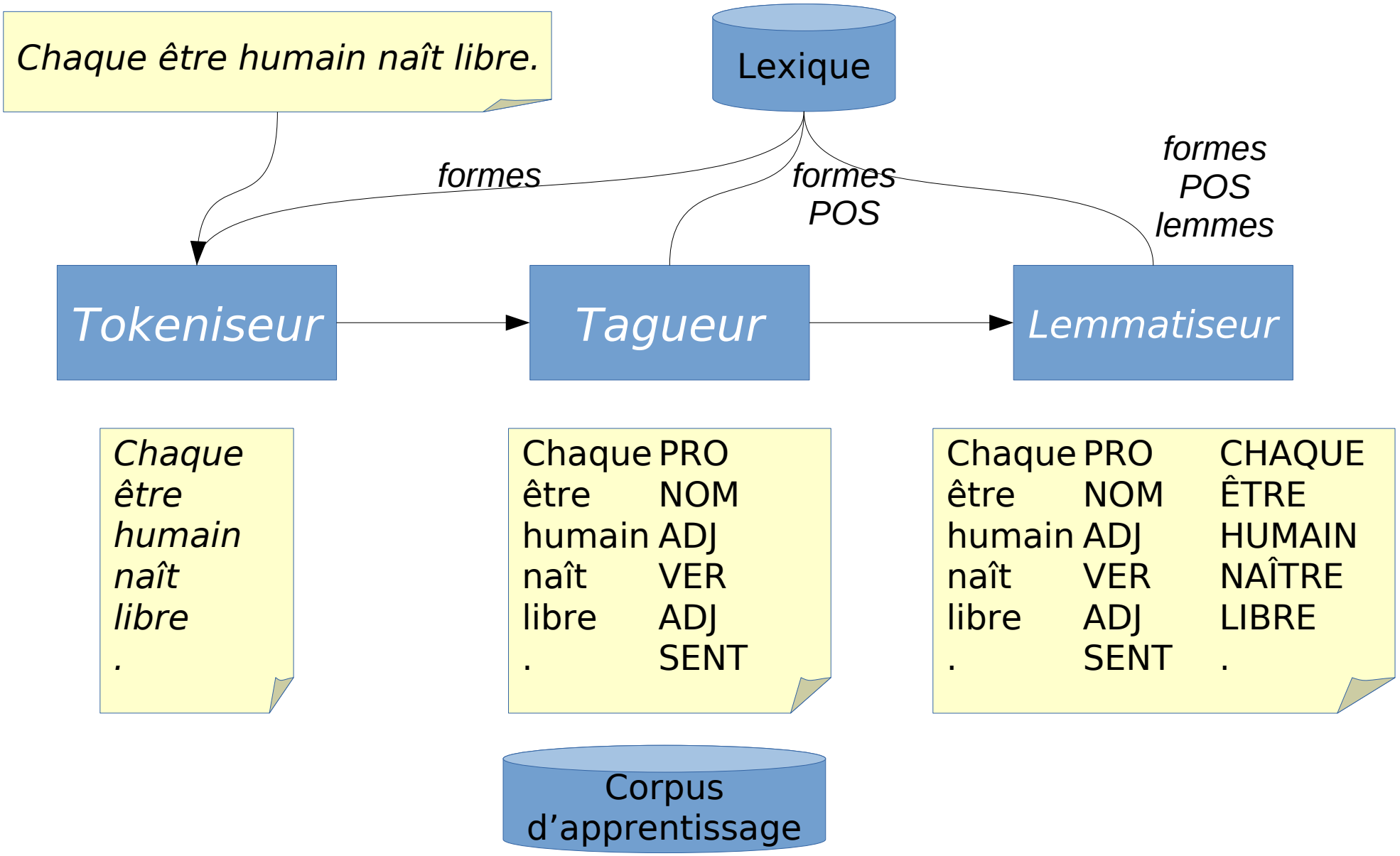
Chaque
être
humain
naît
libre
.

Chaque	PRO
être	NOM
humain	ADJ
naît	VER
libre	ADJ
.	SENT

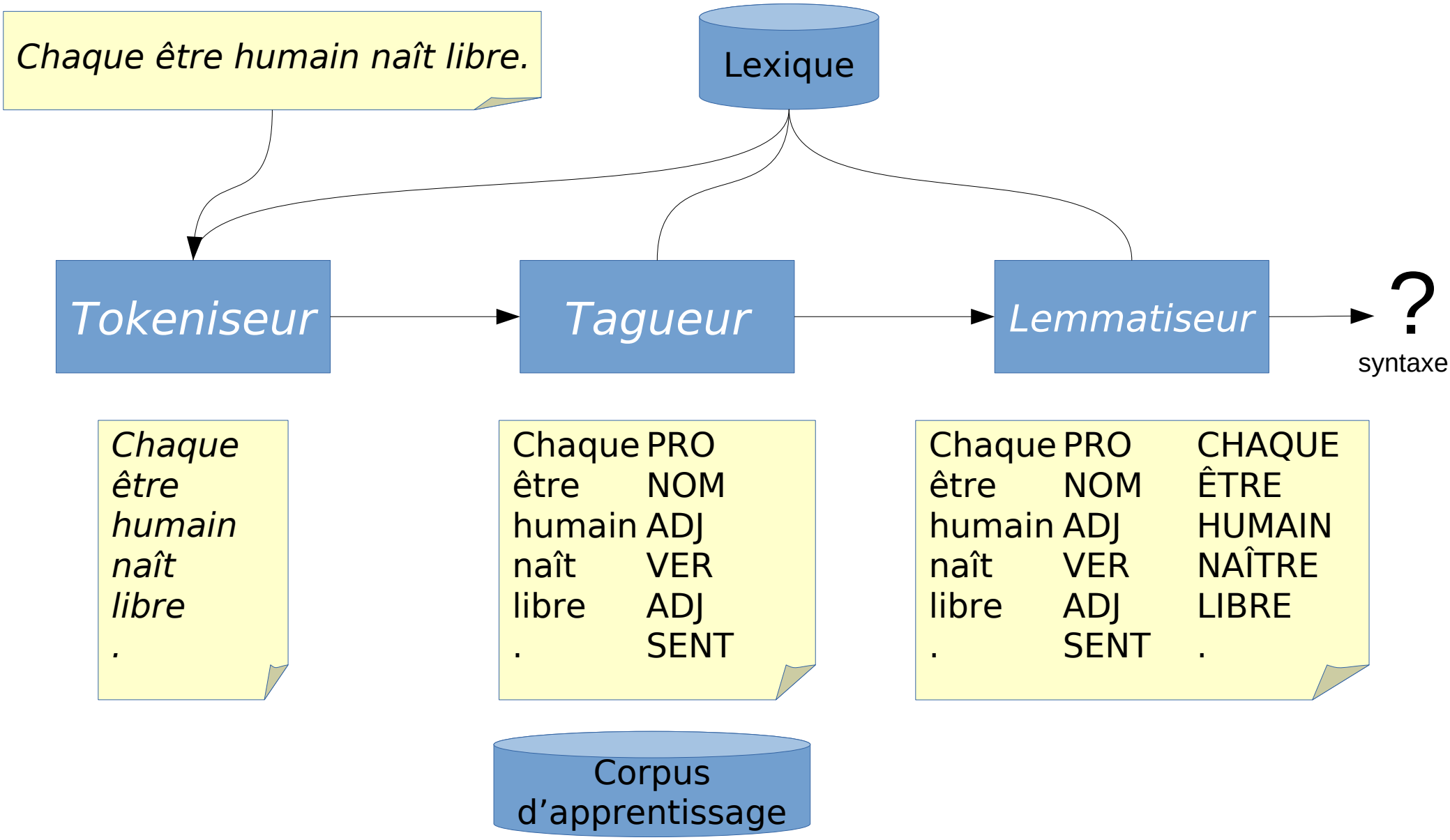
Corpus d'apprentissage

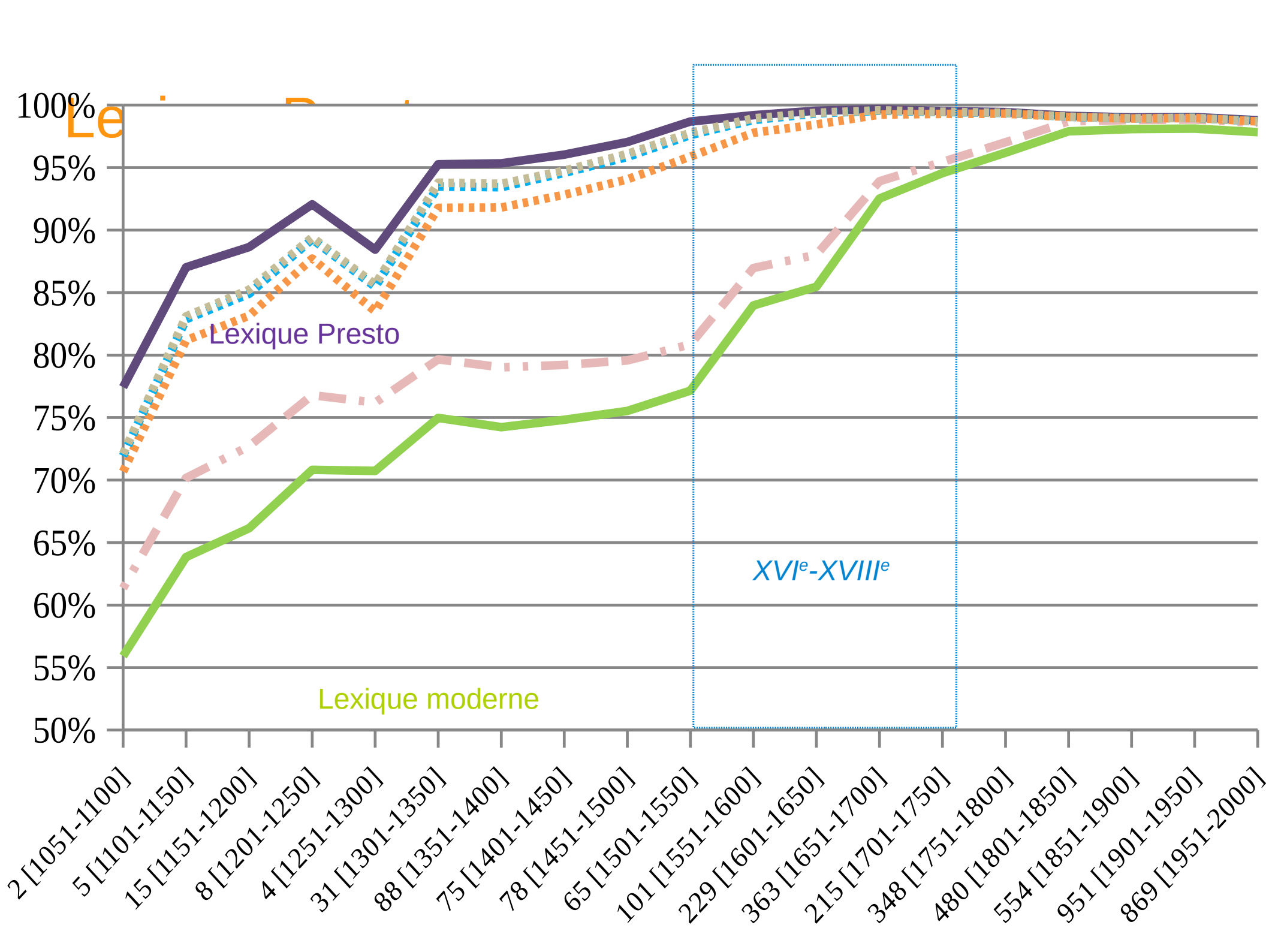


Chaîne de traitement Presto



Chaîne de traitement Presto




100%
95%
90%
85%
80%
75%
70%
65%
60%
55%
50%
Lexique Presto
Lexique moderne
XVIe-XVIIIe
2 [1051-1100]
5 [1101-1150]
15 [1151-1200]
8 [1201-1250]
4 [1251-1300]
31 [1301-1350]
88 [1351-1400]
75 [1401-1450]
78 [1451-1500]
65 [1501-1550]
101 [1551-1600]
229 [1601-1650]
363 [1651-1700]
215 [1701-1750]
348 [1751-1800]
480 [1801-1850]
554 [1851-1900]
951 [1901-1950]
869 [1951-2000]

Analyse syntaxique de l'ancien français

Hops est un analyseur « deep learning » qui combine

- Un modèle de langue général de la famille BERT
- Un analyseur syntaxique « bilinéaire »
- Un étiqueteur morphosyntaxique

Représenter des mots par des vecteurs

motivations et intuitions

- L'hypothèse distributionnelle nous dit que deux mots qui apparaissent souvent dans les mêmes contextes en corpus tendent à avoir un sens similaire
- Alternativement on peut souvent deviner le sens d'un mot sachant le contexte :
  - He filled the **wampimuk**, passed it around and we all drunk some
  - We found a little, hairy **wampimuk** sleeping behind the tree

Vecteurs et similarité

- Un vecteur est une colonne de nombres, en voici deux :

$$\mathrm{x} = \begin{bmatrix} 0.3 \\ 1 \\ 0 \end{bmatrix} \qquad \mathrm{y} = \begin{bmatrix} 0 \\ -0.2 \\ 0.99 \end{bmatrix}$$

- On peut définir une mesure de similarité $s \in [-1, 1]$ entre deux vecteurs

$$s = \cos(x, y) = \frac{\mathrm{x}^T \mathrm{y}}{||\mathrm{x}||||\mathrm{y}||}$$

- Deux vecteurs qui se ressemblent auront une similarité proche de 1 ; deux vecteurs qui ne se ressemblent pas du tout une similarité proche de -1

Word Embeddings

- Les Word embeddings sont des dictionnaires qui associent des chaines de caractères à des vecteurs qui représentent le sens des mots
- Par sens on entend que la mesure de similarité entre vecteurs de mots soit intuitivement cohérente. (synonymes sont proches, non synonymes sont éloignés

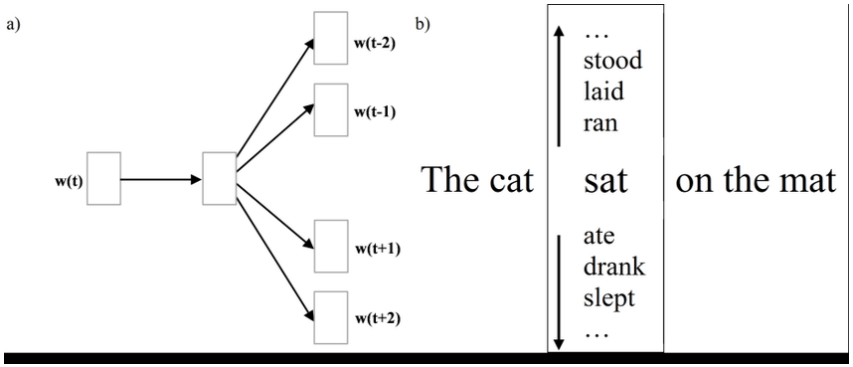
Deux problèmes des word embeddings

- **Pas de gestion de la polysémie et de l'homonymie**. Mélange les vecteurs de avocat_1 et de avocat_2 comme ceux de wampimuk_1 et de wampimuk_2
- **Le vocabulaire n'est pas un ensemble fini** Toujours de nouveaux mots quand on agrandit un corpus, quand on change de domaine. . . (et puis il y a les fautes)

Formalisation de l'hypothèse distributionnelle

Skip-gram

- Les modèles de word embeddings word2vec formalisent l'hypothèse distributionnelle

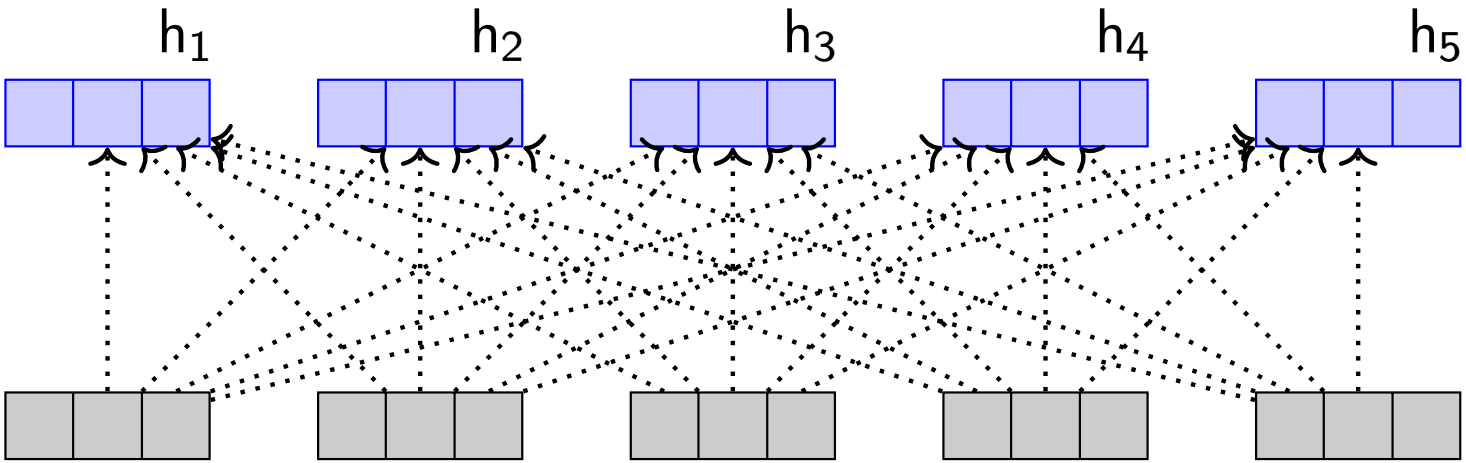


- Un modèle word2vec cherche à apprendre un dictionnaire :
  - <string> $\Rightarrow$ vecteur

  tel que le vecteur permette de prédire le mieux possible les mots du contexte.
- Il en résulte que les vecteurs de mots apparaissant dans les mêmes contextes sont similaires

Modèles contextualisés

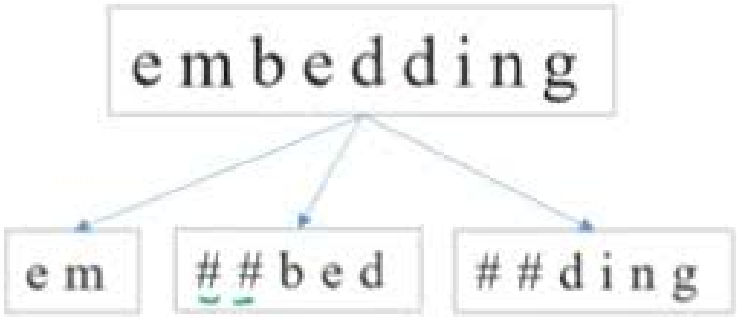
Les modèles contextualisés (transformers) permettent de faire passer des messages entre embeddings :



D'une certaine manière, cela permet de lutter contre le problème de mélange des sens dans un même vecteur

Word pieces

- Même si ce n'est pas obligatoire, la plupart des modèles lexicalisés utilisent des vocabulaires finis.



- Le vocabulaire est construit à partir de caractères de base puis en utilisant les n-grammes les plus fréquents
  - Aucune connaissance en morphologie n'est injectée
- Pour construire l'embedding d'un mot donné, on combine les embeddings de ses morceaux.

Entrainer un modèle de langue contextualisé

Modèle transformer de type BERT

- Un modèle de langue s'entraine en utilisant une méthode qui consiste à masquer un mot et à demander au modèle de prédire le ce mot.
- Rappelons nous le wampimuk :
  - He filled the **wampimuk**, passed it around and we all drunk some
  - He filled the **MASK**, passed it around and we all drunk some
- On compare la prédiction avec le mot **démasqué**. Si celle-ci est fausse l'algorithme d'entrainement change les poids du modèle pour qu'il y ait plus de chances de prédire correctement la prochaine fois. (si la prédiction est juste il peut se renforcer aussi)
- Et on recommence un très grand nombre de fois sur un très gros volume de données (plusieurs dizaines de Gb. Le modèle FlauBERT pour le français est entrainé sur 75Gb de texte pendant plusieurs jours)

Résumé temporaire

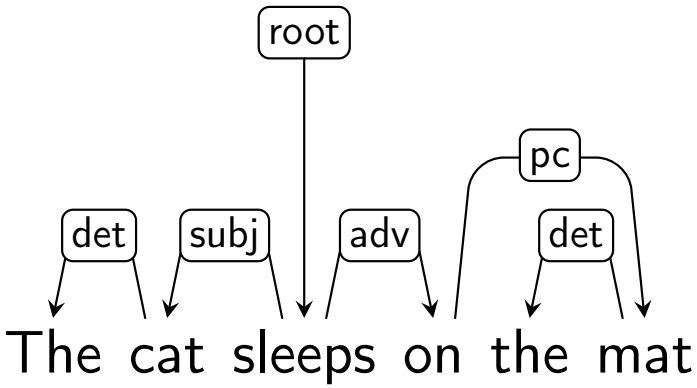
- Un modèle de langue transformer fonctionne comme suit :
  1. Texte segmenté en chunks en utilisant une méthode statistique (em ##bed ##ding)
  2. Association des chunks à leurs embeddings statiques (word2vec)
  3. Transformations des embeddings statiques en embeddings contextualisés (BERT)

Pour l'analyse syntaxique

En pratique on refusionne les vecteurs des sous mots en vecteurs de mots pour l'analyse syntaxique

Analyse syntaxique

- Le but de l'analyse syntaxique est de produire des arbres en dépendances :

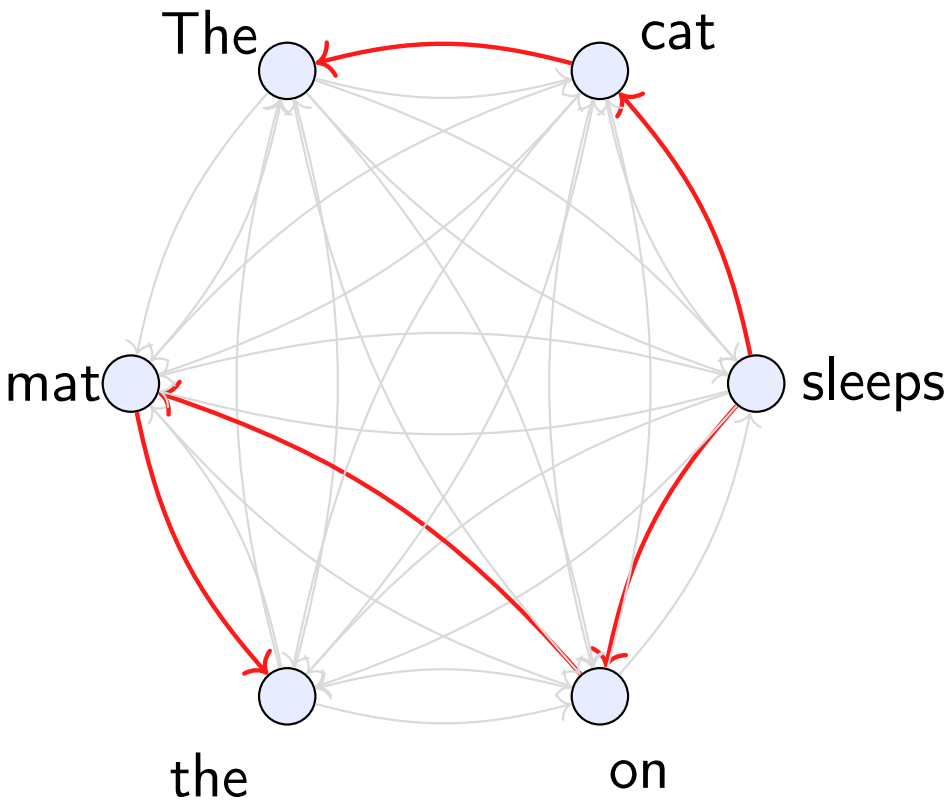


3 étapes principales

1. Obtention d'embeddings contextualisés pour les mots
   - On renforce la linéarisation des mots avec un Bi-LSTM
   - On utilise une méthode pour casser les effets de symmétrie (emb gouverneur, emb dépendant)
2. Prédiction d'un arbre
3. Prédiction des relations

Prédiction de l'arbre

- Prédiction d'un score (nombre réel) pour tous les arcs possibles, $\text{score}(x, y) = x^T W y$
- Utilisation d'un algorithme qui trouve l'arbre maximal couvrant d'un graphe complet orienté (Chu Liu Edmonds) :

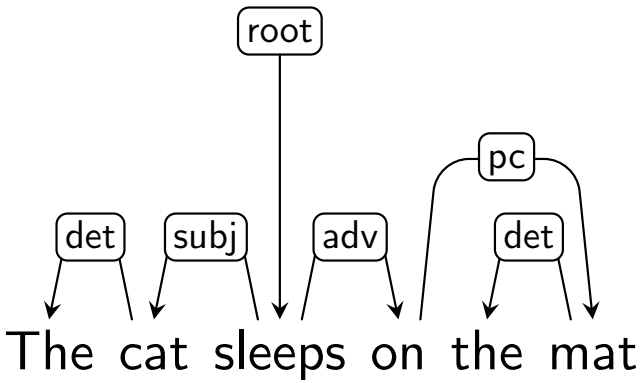


Prédiction des relations

- Pour chaque arc de l'arbre couvrant,
  - Score de chaque étiquette l :

$$score(x, y, l) = x^T W_l y$$

    dont l'étiquette avec le score maximal est choisie comme étiquette de l'arc.
  - Si tout se passe bien



Mode opératoire

Le modèle de langue est entrainé sur un gros corpus (dizaine de gigas). Le parser est entrainé sur un tout petit corpus : quelques miliers de phrases (ex. Universal Dependencies)

Etiquetage de séquences

- Le modèle a une architecture multi-tâches
- Etant donné un codage des mots par une liste de vecteurs
  - On prédit la structure syntaxique
  - Mais aussi les catégories morphosyntaxiques (pos tags)

Eviter la propagation d'erreurs

Contrairement aux architectures classiques en pipeline : prédiction des tags suivie de prédiction de l'arbre de syntaxe, ici la prédiction de l'arbre de syntaxe n'est pas directement dépendante de la prédiction des tags

Compte rendu sur l'ancien français

Projet Profiteroles

- **But :** produire des analyses syntaxiques pour l'ancien français
- **Problème** Données brutes limitées :

Corpus	**Size (Mb)**	**Size (Mwords)**
BFM	20.7	3.91
AND	17.2	3.25
NCA	9.7	2.05
Chartes Douai	3.1	0.56
OpenMedFr	1.7	0.33
Geste	1.5	0.32
MCVF	1.4	0.26
Chartes Aube	0.2	0.04
Total	55.3	10.53

- Données annotées (SRCMF) : $\approx$ 18000 phrases ; 40000 mots
- Absence de norme dans la langue (variations régionales, stylistiques etc.)

Méthodes

- Baseline I (word2vec seulement, sans modèles contextuels)
- Baseline II modèles existants (français moderne ou multilingues)
- En entrainant un modèle de langue sur l'ancien français
- En affinant un modèle préexistant sur l'ancien français

Baselines I et II

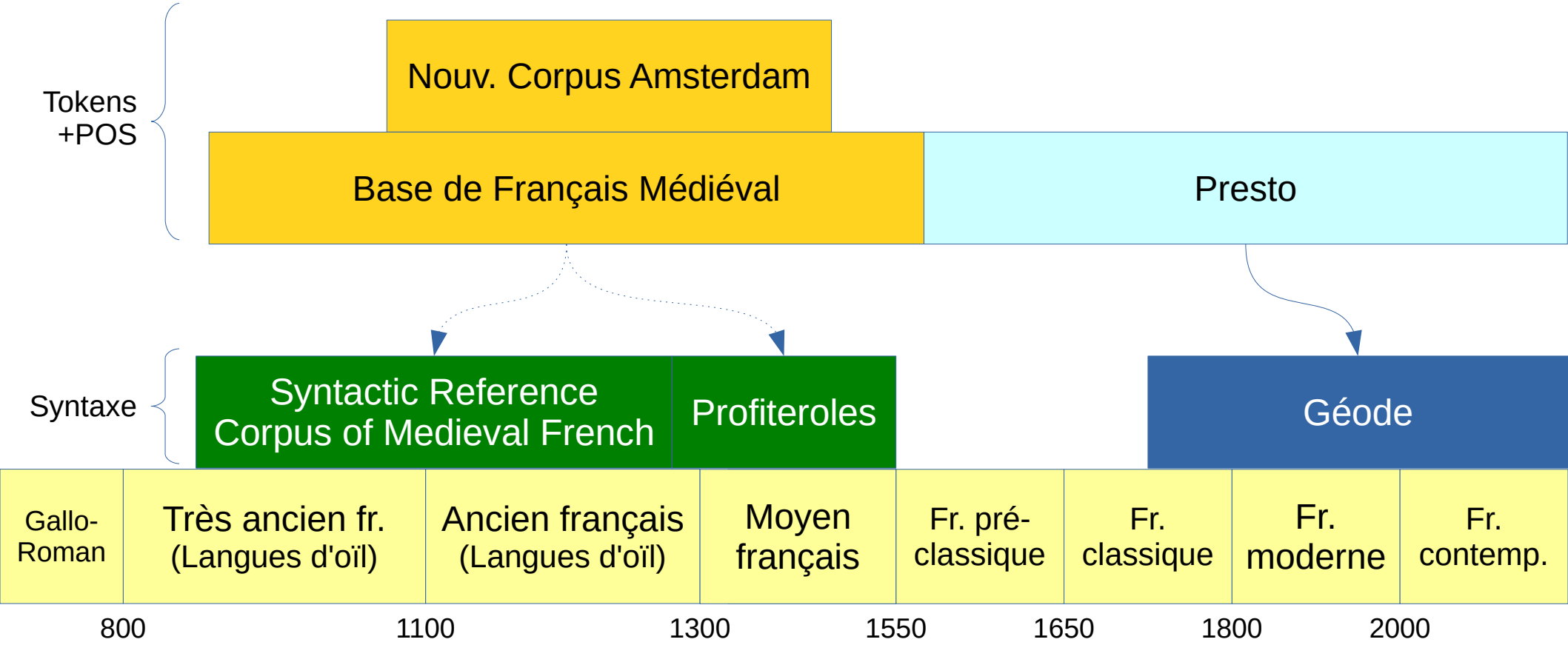
Embeddings	**UPOS**	**UAS**	**LAS**
Vanilla	93.51	87.60	81.54
FlauBERT	95.70	90.43	85.45
CamemBERT	95.86	91.15	86.31
mBERT	96.06	91.52	86.83

Dans tous les cas les modèles (de langue) ne sont pas entrainés sur des données Ancien Français

Modèles de langue ancien français

BERTrade-scratch	96.74	92.37	88.42
BERTrade-mBERT	96.95	93.33	89.60
BERTrade-CamemBERT	97.16	93.75	90.06
BERTrade-FlauBERT	96.94	93.75	90.07

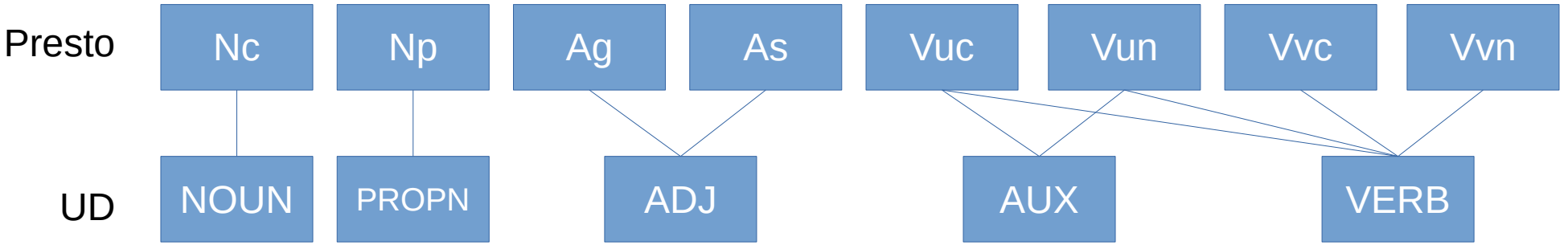
Projet Géode
2020-2024



Projets en cours pour ajouter des étages syntaxiques

Projet Géode
2020-2024

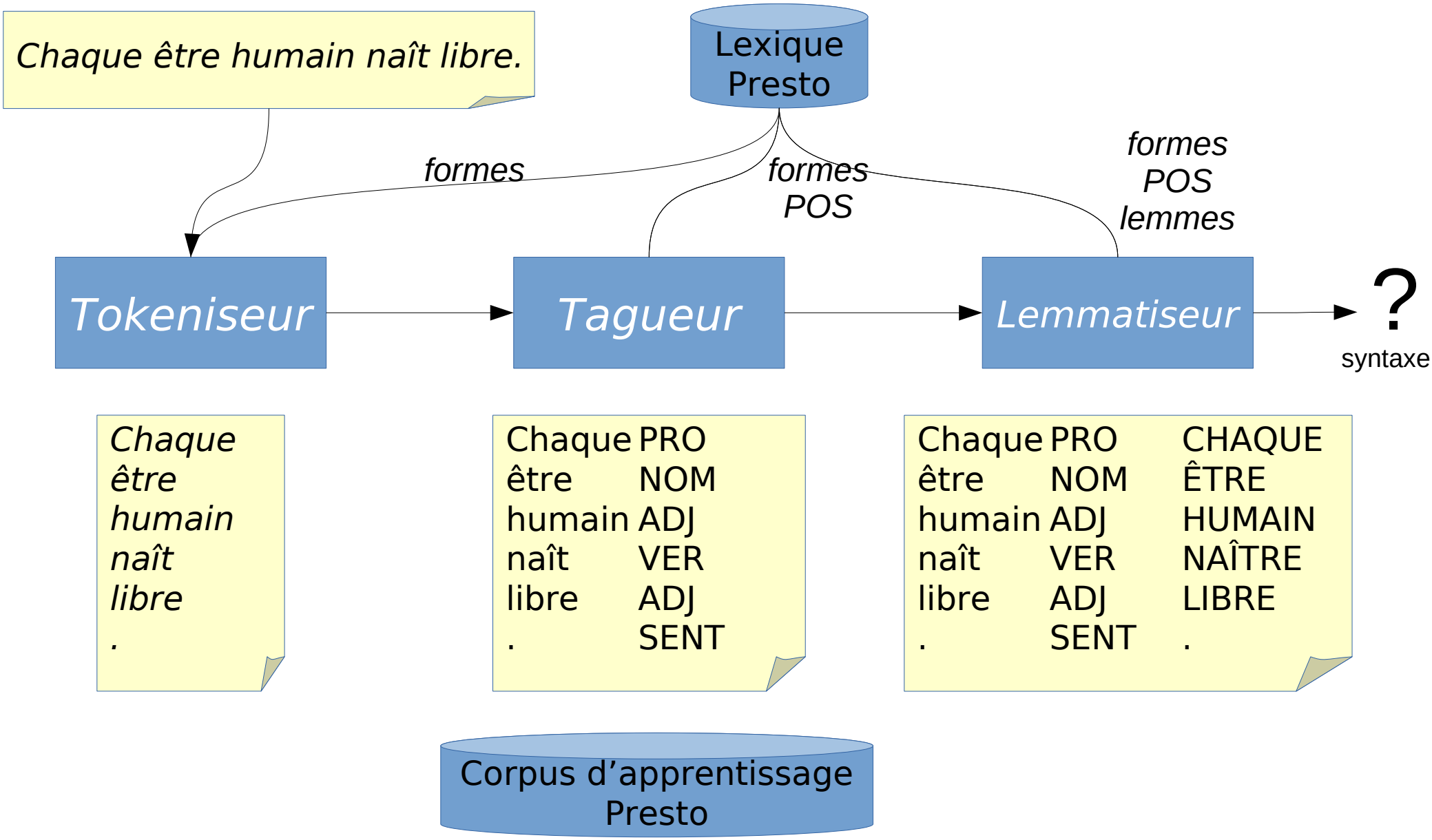
- Choix d'un tagset
  - Presto (2013) : à l'époque, pas de tagset dominant, création d'un tagset *adhoc* à partir de tagsets existants
  - Géode (2020) : *Universal dependencies (UD)*
- Alignement des tagsets



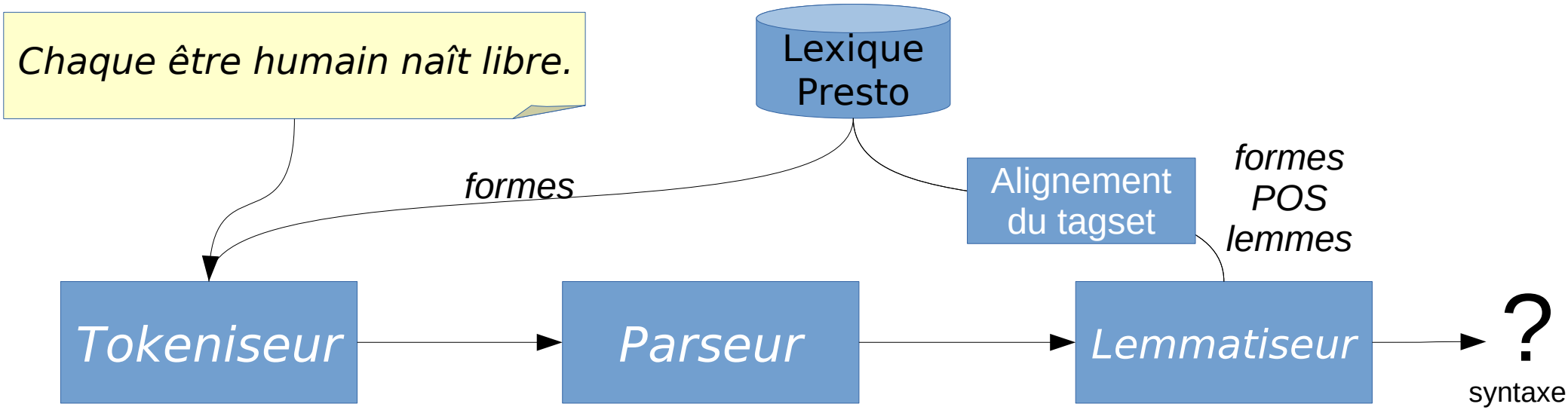
Projet Géode
2020-2024

- Choix d’un corpus d’apprentissage adapté
  - Pas de corpus arboré du français classique
  - Recherche d’un corpus du français moderne
    - Utilisable pour du français classique
    - Tokenisation ~compatible avec Presto
  - => SEQUOIA

Chaîne de traitement Presto



Chaîne de traitement Géode



1 Chaque
2 être
3 humain
4 naît
5 libre
6 .

1	Chaque	PRO	2	det
2	être	NOM	4	nsubj
3	humain	ADJ	2	amod
4	naît	VER	0	root
5	libre	ADJ	4	xcomp
6	.	SENT	4	punct

1	Chaque	PRO	2	det	CHAQUE
2	être	NOM	4	nsubj	ÊTRE
3	humain	ADJ	2	amod	HUMAIN
4	naît	VER	0	root	NAÎTRE
5	libre	ADJ	4	xcomp	LIBRE
6	.	SENT	4	punct	.

Corpus d'apprentissage SEQUOIA

Évaluation

- Création d’un corpus de référence
  - Pour l’évaluation
  - À la main !
  - Problème : guide d’annotation

Évaluation

- Création d'un corpus de référence
  - ConllUEditor : https://github.com/Orange-OpenSource/conllueditor