Lemmatisation et classification sémantique dans un corpus latin en diachronie longue

L'exemple des sèmes sexuels

Thibault Clérice
clerice.thibault@algorythme.net
@ponteineptique

PSL-ENS, LATTICE

24 Février 2023

Sommaire

Qu'est-ce que le latin ?

1. Une langue dont les textes les plus anciens qui nous sont parvenus viennent majoritairement du **3 BCE** (mais rares inscriptions remontant jusqu'au 6 BCE)
2. Une langue encore en **utilisation aujourd'hui** par certaines administrations (*ie* le Vatican, les diplômes de certaines universités, etc.) et une petite communauté de latin parlé (langue non maternelle)
3. Une langue d'**inscriptions**: une partie non négligeable du corpus antique et médiévale se trouve sur les murs et objets.
4. Une langue **administrative** pour une grande partie de l'Europe occidentale voire orientale pendant le Moyen-Âge
5. Une langue **littéraire** (néo-latin), exemple des auteurs croates écrivent en latin du 15e au 19e.
6. Une langue de **recherche**: Jean Jaurès a ainsi écrit sa thèse en latin en 1891: `https://www.hs-augsburg.de/~harsch/Chronologia/Lspost19/Jaures/jau_soc0.html`.

Et qu'est-ce que le latin antique ?

Sur postulat Antiquité = 3 BCE – 5 CE

1. Une langue dont la majorité des **témoins d'époque** sont des **inscriptions**, des **graffitis** (Pompéi) et des **papyri** (Égypte).
2. La grande majorité des textes littéraires nous sont connues par des **copies datant de plusieurs siècles après leur rédaction**. Voire, certains textes ne nous sont connus que par des manuscrits très récents: les *Priapées* (Anonyme) datent du 1 CE mais le plus ancien manuscrit les contenant est un manuscrit de Boccace (14 CE): 13 siècles séparent la rédaction originale de la copie.
3. Les copies sont donc majoritairement des copies médiévales (8 CE et après).
4. Certains textes, en particulier les commentaires de grammairiens, ont probablement perdu intégralement leur structure (*cf.* le projet HyperDonat `https://www.youtube.com/watch?v=E6W6h4Qhve8`)

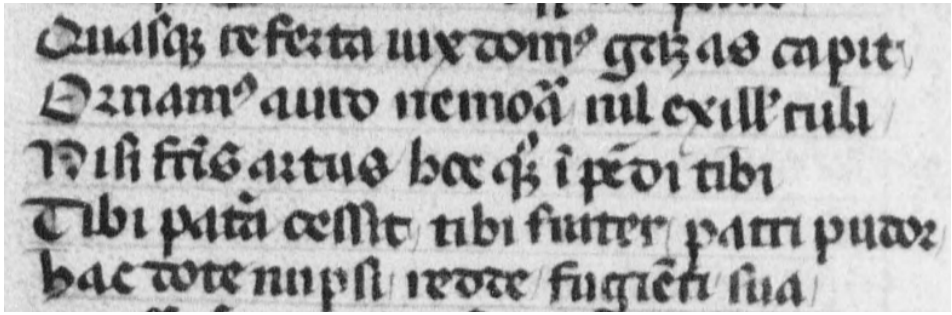
Que lit-on quand on lit aujourd'hui du latin ?

Les étudiant-e-s ne lisent que des documents édités par des chercheurs qui sont généralement:

1. **Normalisés** en terme de graphie: on reconstitue une "orthographe" originale cicéronienne,
2. Issus d'une **sélection de leçons** entre les différentes versions du texte circulant sur les manuscrits,
3. **Structurés** d'après des découpes parfois médiévale ou plus récentes (paragraphes, phrases, etc.).

La philologie est entre autre la science qui régit ces pratiques. Le latin partage certains de ses traits avec d'autres langues, notamment celle de la variation graphique au moyen-âge avec les langues vernaculaires telles que l'ancien français.

Médée: Manuscrit vs. Édition



Quasq; referta uix domꝰ gazas capit
Ornamꝰ auro nemoã nil exilł tuli
Nisi frīs artus hoc q̊; ī pēdi tibi
Tibi patà cessit tibi frater patri pudor
hac dote nupsi redde fugiẽti sua

quas quia referta uix domus gazas capit, 485
ornamus auro nemora, nil exul tuli
nisi fratris artus : hos quoque impendi tibi ;
tibi patria cessit, tibi pater, frater, pudor : —
hac dote nupsi ! — Redde fugienti sua.

472 som(p)noque *A* : summoque *E* || **475** iussasque *ω* : ausasque *Heinsius Zwierlein* || **477** *post* **482** *transp. Leo post* **487** *Delrius del. Zwierlein* || quaerens *E* : sequens *A* || **484** agunt *E* : petunt *A* || **485** quas quia *E* : quasque *E* || gazas *E* : gazis *uel* gaza *Bentley* || **488** tibi pater frater *Ascensius* : tibi pater. tibi frater *E* tibi frater pater *A* || **492** munus *E* : mitius *A*.

Médée 485-489, Sénèque.
À gauche, manuscrit *BnF Latin 6395* microfilmé et transcrit pour CREMMA.
À droite, édition Budé - Les Belles Lettres, établie par Chaumartin, impression 2008, p. 175.

Corpus fac-similaire d'éditions

Les éditions scientifiques reproduisant des textes souvent deux fois millénaires, les droits d'auteurs sont tombés.
Les premiers projets de corpus numériques remontent aux années 1960-1970 pour le Grec ("Lettres Classiques") puis fin des années 1980 et début des années 1990 pour le latin. Certains corpus sont nativement ouvert ou à prix coûtant (Perseus, PHI [Packard]), tandis que d'autres sont directement commerciaux (*Patrologia Latina*, *CLCLT*).
Aujourd'hui, trois grands corpus ou projets de données ouverts pour le latin classique ($<$ 2 CE) et tardif ($<=$ 5 CE):

- *Perseus*, et le projet "enfant" *Open Greek and Latin*,
- *DigilibLT* pour les textes tardifs non-chrétiens,
- le *Corpus Corporum* qui est un agrégat de corpus.

À côté de ces corpus, il existe des corpus thématiques: autour d'un auteur, d'un genre (grammairiens par exemple), d'une époque, d'un sujet (droit). Et quelques (très très) rares éditions numériques.

Sommaire

Rappel rapide sur le système morphologique du latin

- 6 cas (7 en incluant le locatif pour certains mots)
- 2 nombres (singulier, pluriel)
- 3 genres (masculin, neutre, féminin)
- 3 personnes (1, 2, 3) (similaire au Français)
- 7 modes verbaux (Indicatif, Impératif, Subjonctif, Infinitif, Participe, Adjectif, Gérondif, Supin)
- 6 temps (présent, imparfait, parfait, plus-que-parfait, futur, futur antérieur)
- 4 "voies" (Actif, Passif, Déponent, Semi-déponent)
- 3 degrés (Positif, Comparatif, Superlatif)

\+ l'absence de la marque

État des lieux des corpus classiques et tardifs

	Tokens	Ponctuation comprise	Nombre d'auteurs	Nombre de textes	Lemmes uniques	Référentiel
PROIEL	225 064	Non	5	6	7 246	Lewis[1]
Perseus	79 670	Oui	12	12	6 017	Lewis
Harrington	120 029	Oui	9	12	7 675	Lewis
LASLA	**1 630 825**	Non	**18**	**100+**	25 135	Forcellini

Table: Résumé des informations sur les quatre corpus disponibles. Il existe 137 œuvres au sens du LASLA, mais certaines sont des des découpes inhabituelles, nous préférons donc la notation 100+ ici.

État des lieux des corpus classiques et tardifs

	Tokens	Ponctuation comprise	Nombre d'auteurs	Nombre de textes	Lemmes uniques	Référentiel
PROIEL	225 064	Non	5	6	7 246	Lewis[1]
Perseus	79 670	Oui	12	12	6 017	Lewis
Harrington	120 029	Oui	9	12	7 675	Lewis
LASLA	**1 630 825**	Non	**18**	**100+**	25 135	Forcellini

Table: Résumé des informations sur les quatre corpus disponibles. Il existe 137 œuvres au sens du LASLA, mais certaines sont des des découpes inhabituelles, nous préférons donc la notation 100+ ici.

Problème, diffèrent:

1. Les **référentiels POS** et morphologiques
2. Les **lexicalisations** prises en compte, y compris quand le même référentiel lemme est utilisé.

Après le 6e siècle

D'autres corpus tardifs et médiévaux existent, notamment via *Universal Dependencies*:

- UD-LLCT (Late Latin Charter Treebank)
- UD-Universal Dante
- UD-ITTB (Index Thomisticus Treebank)
- PaLaFra (`https://www.palafra.org`)

mais présentent les mêmes problèmes du point de vue des lemmes ou des pratiques d'annotations.
LiLa (Linking Latin) comme projet de base de connaissances connectant les référentiels (mais ne résout pas l'ensemble des problèmes):
`https://lila-erc.eu/`.

Représentativité Temporelle du Corpus

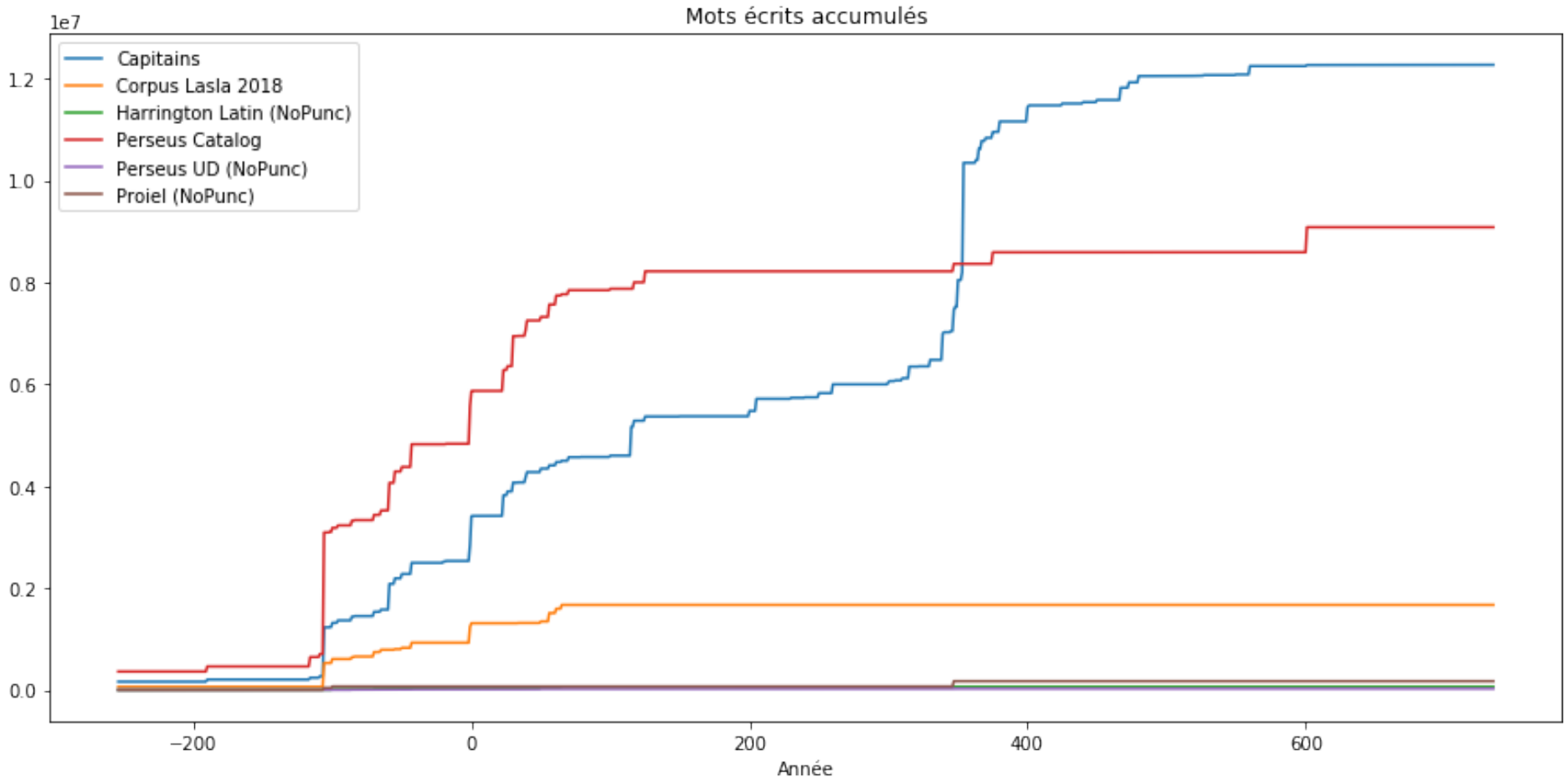


Figure: Mots accumulés (Année de naissance de l'auteur)

Problèmes avec un corpus de linguistes

Forme	Mode	Temps
amatus (sum)	Indicatif	Parfait
amatus (eram)	Indicatif	Plus-que-parfait
amatus (ero)	Indicatif	Futur antérieur
amatus (sim)	Subjonctif	Parfait
amatus (essem)	Subjonctif	Plus-que-parfait
amatus (esse)	Infinitif	Parfait
amatum (iri)	Infinitif	Futur

Table: Annotations possibles pour la forme *amatus* dans le LASLA, hors participes

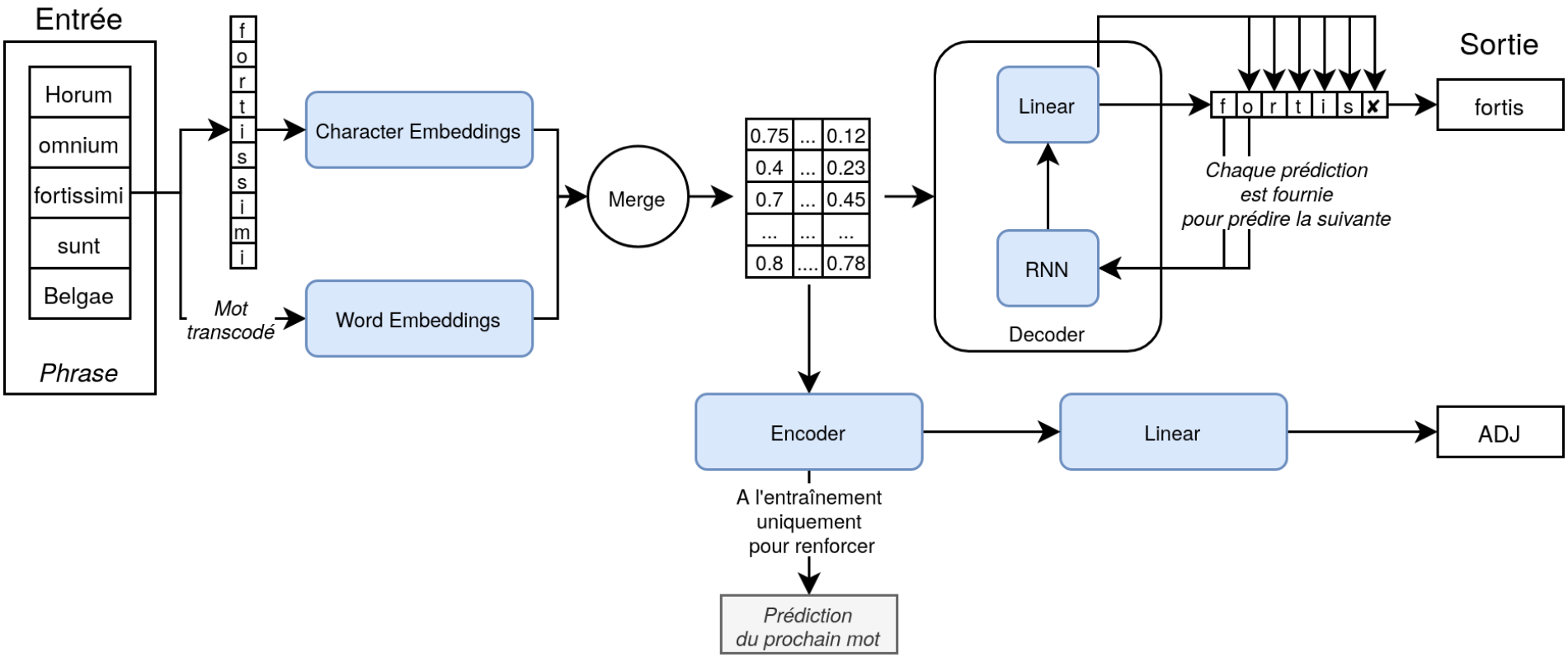
Participe qui porte le temps de l'auxiliaire en ellipse.
Autre problèmes: formes de lemme discutables (PAEDICO → PEDICO), désambiguïsation tantôt liée à l'**homographie des formes canoniques** mais à la **morphologie différente**, au **genre** différent (vallus1 / vallus2) ou aux **fonctions grammaticales** différentes (qui1, qui2, qui3, etc.) mais quelque fois aussi **sémantique**.

Deux corpus nouveaux

- **Échantillonnage** *Du IIème siècle à Thomas More, un corpus gold de latin lemmatisé et annoté en morpho-syntaxe*, A. Glaise et T. Clérice, [doi]10.5281/zenodo.6383162. 57.000 tokens, 40 textes (généralement 500 à 600 tokens), principalement 2 CE - 10CE puis 20.000 tokens Jacques de Voragine et Thomas More.
- **Domaine** *Lemmatisation et analyse morpho-syntaxique des Priapées*, T. Clérice, https://github.com/lascivaroma/priapea-lemmatization.

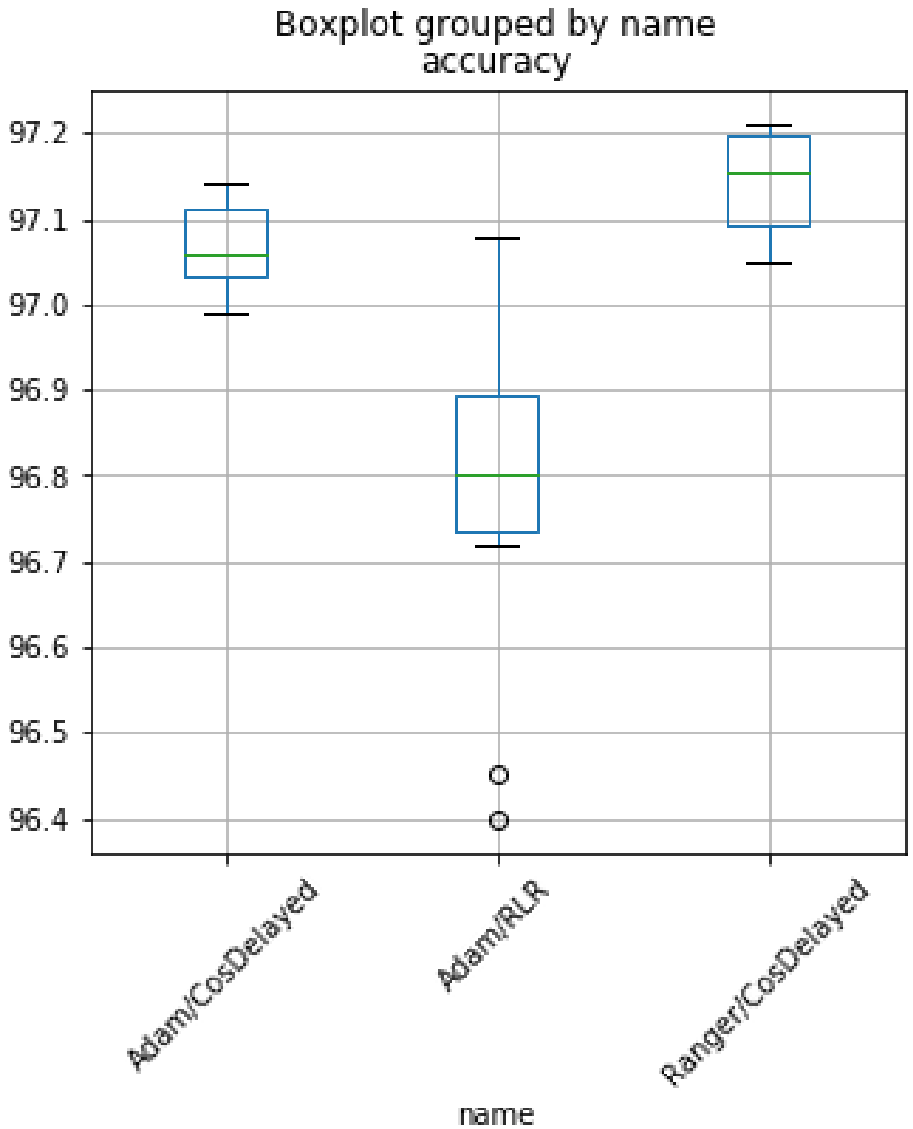
Pie

Improving Lemmatization of Non-Standard Languages with Joint Learning, E. Manjavacas, Á. Kádár et M. Kestemont, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Doi: 10.18653/v1/N19-1153.



PaPie

- Ajout d'optimisations pour l'inférence,
- Ajout de bruit aléatoire (par exemple, capitalisation d'un mot ou de la phrase),
- Optimiseur et LR (`https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer`),
- Autres cible en score pour le dev.



Modèle 1: LASLA (tout court)

	Accuracy		
	Formes connues	Formes Inconnues	Formes Concernées
Lemme	97,41	92,92	
POS	96,49	92,45	
Genre	96,28	91,49	89,98
Nombre	97,02	93,85	96,44
Cas	92,34	87,37	87,84
Degré	98,07	93,96	93,37
Mode Temps Voix	98,35	90,80	94,44
Personne	99,71	98,15	98,49
Tâches agrégées	85,86	76,68	

Table: Résultats du modèle final. Les formes concernées sont les formes dont le trait morphologique n'est pas absent: par exemple, l'*accuracy* de 87,84% en cas correspond au taux de succès pour toutes les formes qui sont annotés pour ce trait morphologique, excluant les formes verbales non participiales, les prépositions, etc.

Tests sur *Glaise* et *Priapées*

Catégorie	*Accuracy*		*Accuracy* quand applicable	
	Priapées	Tardif	Priapées	Tardif
lemma	94,2	94,5	N/A	N/A
Deg	96,8	97,5	90,3	91,5
Numb	94,7	96,5	94,1	96,4
Person	99,0	99,7	96,0	99,1
Mood_Tense_Voice	96,1	97,7	87,0	92,3
Case	89,4	92,9	84,1	88,0
Gend	91,7	92,8	77,8	79,2
pos	95,0	67,6	N/A	N/A

Table: Résultat du modèle sélectionné sur le corpus des *priapées* et de latin tardif. L'*accuracy* quand applicable désigne la précision de l'algorithme sur les termes qui nécessitent une annotation, en d'autres termes, pour la personne par exemple, seules les annotations sur les verbes sont concernées.

Étude des erreurs

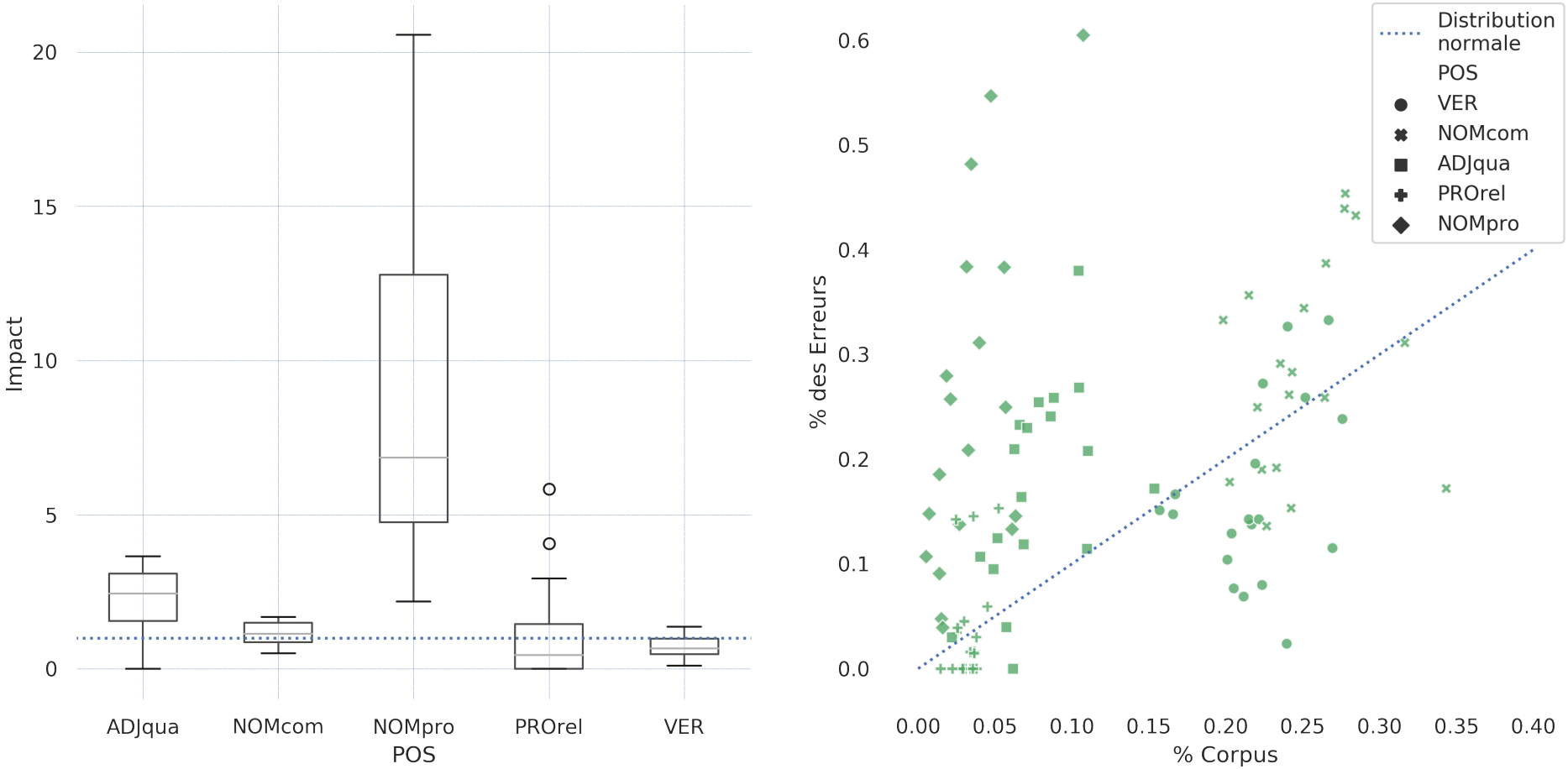


Figure: Responsabilité des POS dans les erreurs du lemmatiseur sur le corpus tardif.

LASLA+

	Tous			Tokens connus			Tokens Inconnus			Tokens Ambigus		
Tâche	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec
Cas	94,64	90,38	88,82	94,86	91,05	89,56	90,18	74,51	70,96	88,53	87,46	86,19
Degré	98,41	97,61	97,83	98,59	97,78	98,1	94,72	94,2	93,23	93,57	93,78	94,69
Genre	97,44	94,29	93,99	97,63	94,55	94,39	93,5	90,09	88,15	91,71	92,12	92,14
Mode/Temps/Voie	98,9	90,21	86,81	99,07	90,88	88,88	95,37	81,31	83,74	93,91	83,72	82,34
Nombre	98,12	98,1	97,92	98,26	98,24	98,06	95,3	93,68	93,33	93,87	93,83	93,33
Personne	99,77	99,12	98,2	99,83	99,25	98,64	98,49	98,05	95,65	98,29	96,08	93,33
Lemme	97,32	84,41	84,09	97,72	90,53	90,52	89,23	76,31	75,96	92,56	69,62	70,3

Table: Résultats du modèle LASLA+. Acc=*Accuracy*, Pre=*Précision*, Rec=*Recall*.

Le problème des données propres

	A	B	C	D
Normalisation des lettres	Non	Oui	Non	Oui
Suppression des formes inconnues (points)	Non	Non	Non	Oui
Tokenisation Phrase	Ponctuation Forte	Ponctuation Forte	35 Mots	35 Mots
lemma	**-1,51**	-0,22	**-1,62**	-0,23
Deg	-0,30	-0,16	-0,26	-0,06
Numb	-0,41	-0,31	**-0,56**	-0,12
Person	-0,05	0,00	-0,06	-0,03
Mood_Tense_Voice	-0,11	-0,02	-0,12	-0,04
Case	**-0,76**	**-0,66**	**-1,41**	**-0,32**
Gend	-0,28	-0,13	-0,23	-0,11
pos	**-0,57**	-0,33	**-0,66**	-0,18

Table: Impact de la normalisation sur l'*accuracy*, en points de %.

Pie Extended

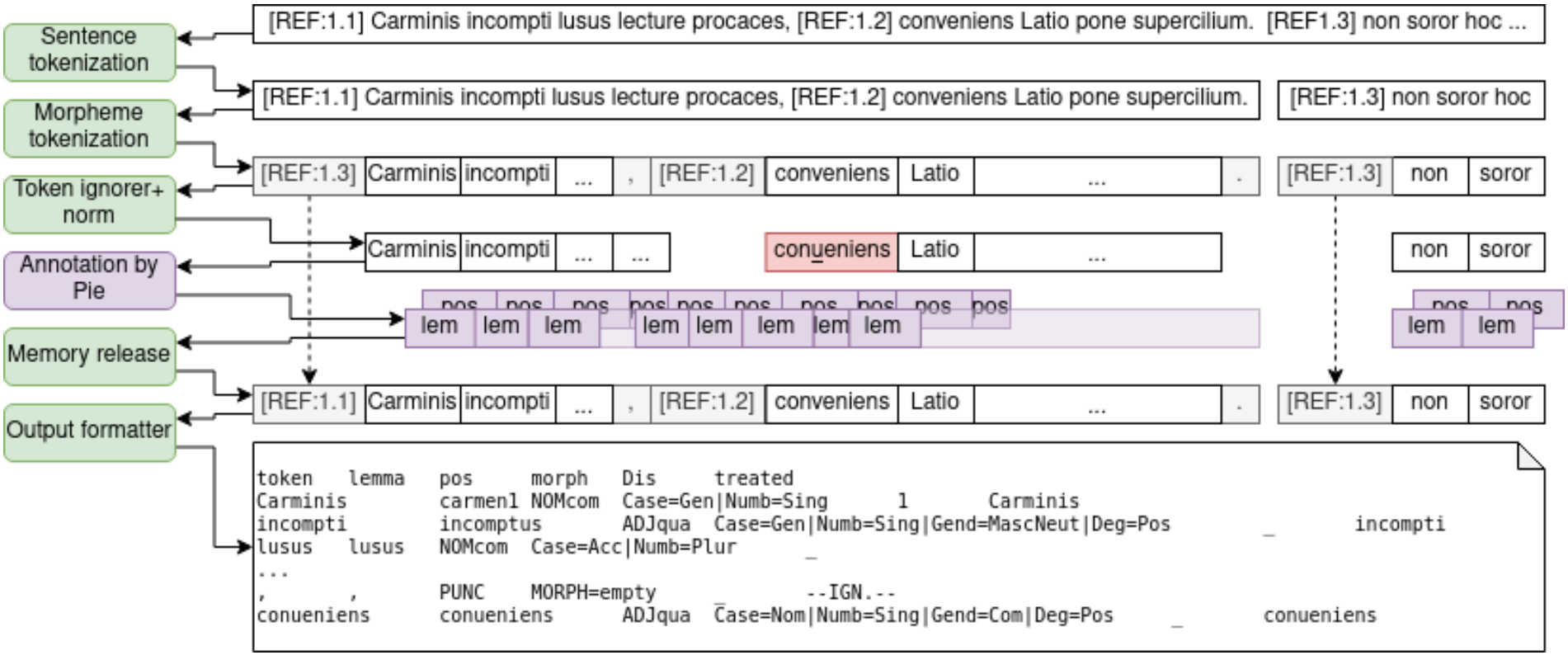


Figure: Fonctionnement de Pie-Extended

Sommaire

Constituer des corpus pour l'étude du lexique en diachronie

- Évolution de certains lexiques intéressant à la fois pour l'évolution de la langue en diachronie et pour l'étude des cultures.
- Constituer une base de données d'occurrences:
  - Facile pour les occurrences des termes connus et non-ambigus,
  - Plus difficile pour les usages ambigus,
  - Encore plus complexe (et coûteux) pour les utilisations figurées, parfois uniques.

Ce que l'on cherche

- Être capable de **détecter les termes courants** (*futuo, pedico, etc.*)
  - "Quid faciat uolt scire Lyris. Quod sobria: fellat.", *Épigrammes* II.73, Martial
  - "Lyris veut savoir ce qu'elle fait [ivre]. Ce qu'elle fait sobre: elle suce." (Traduction D. Noguez, Arléa)
- Être capable de détecter les **usages figurés fréquents**.
  - "Tanta est quae Titio columna pendet , Quantam Lampsaciae colunt puellae.", *Épigrammes* XI.51, Martial
  - "La colonne qui pend à Titius est si grande que les jeunes filles de Lampsaque la vénèrent."
- Être capable de détecter les **usages complexes** (métaphore, analogies, figures plus complexes comme l'adynaton).
  - "Donec proterua nil mei manu carpes, licebit ipsa sis pudicior Vesta.", *Priapées*, 31.
  - "Tant que tu ne cueilles rien chez moi d'une main effrontée, tu pourras être plus pudique que Vesta elle-même."

Constitution d'un exemplier

- Basé sur le *Latin Sexual Vocabulary* d'Adams, à l'exception de l'épigraphie.
- 2516 extraits, où le(s) mot(s) analysé(s) par Adams sont qualifiés et annotés.
- Ajout des informations sur le texte source et de la source bibliographique.

En ligne: `https://github.com/lascivaroma/exemplier`, `https://dev.chartes.psl.eu/lascivaroma`.

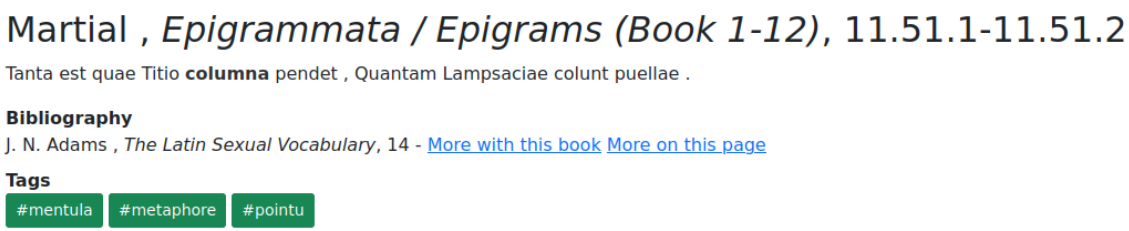


Figure: Exemple de visualisation d'un extrait.

Littérature latine et sexe ?

- Littérature scientifique
  - Plutôt antiquité tardive pour les textes médicaux sur les humaines
  - Souvent des traductions de textes grecs
  - Pour le reste, traite surtout de la sexualité des animaux.
- Poésie “classique” (Principalement érotique / bucolique)
- Poésie satirique (Martial, Ausone, Juvénal, Pétrone... Très aggressif)
- Prose historique, “légale”, commentaires religieux (Principalement diffamatoire ou négatif)

Exemplier

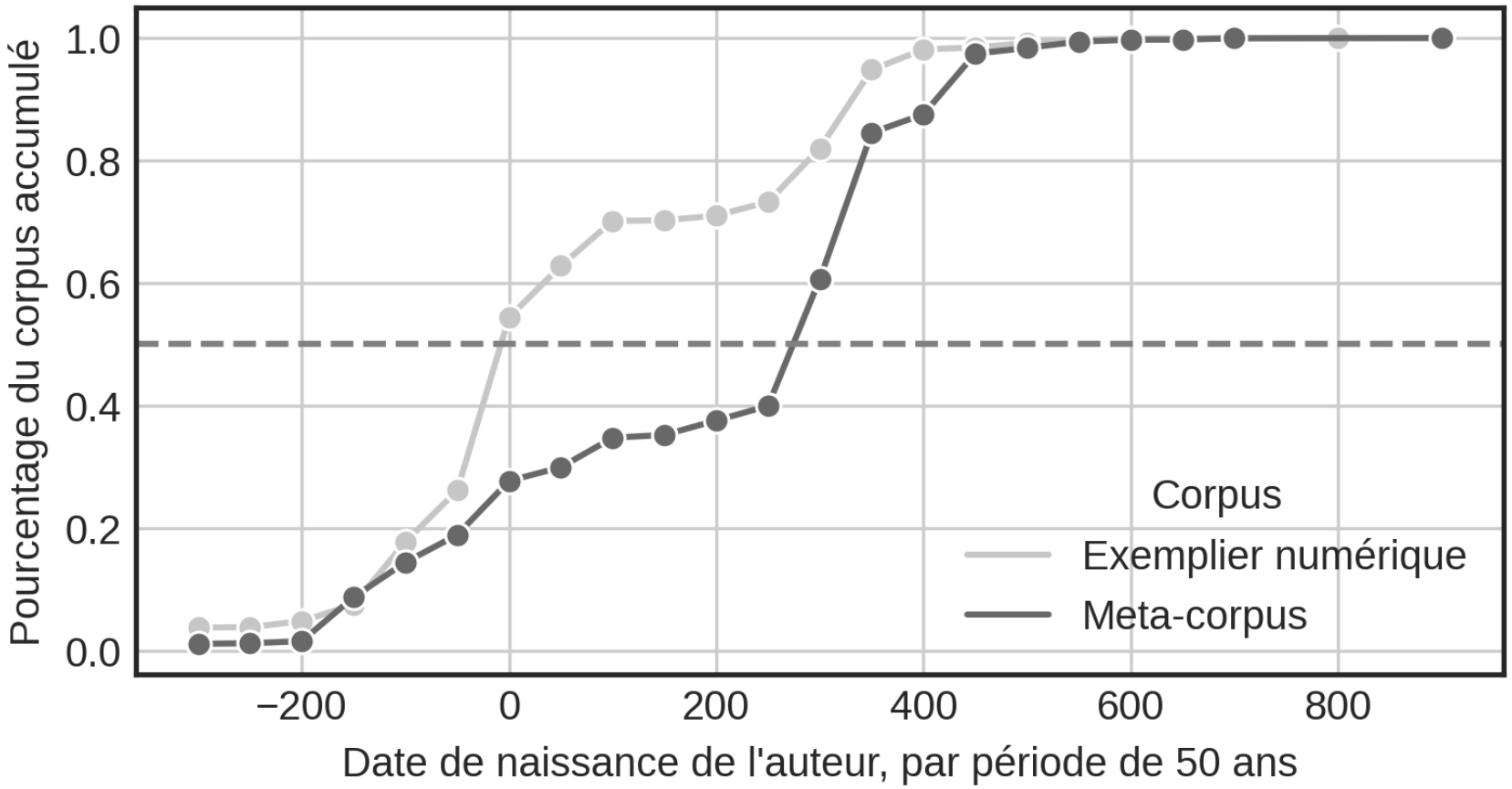
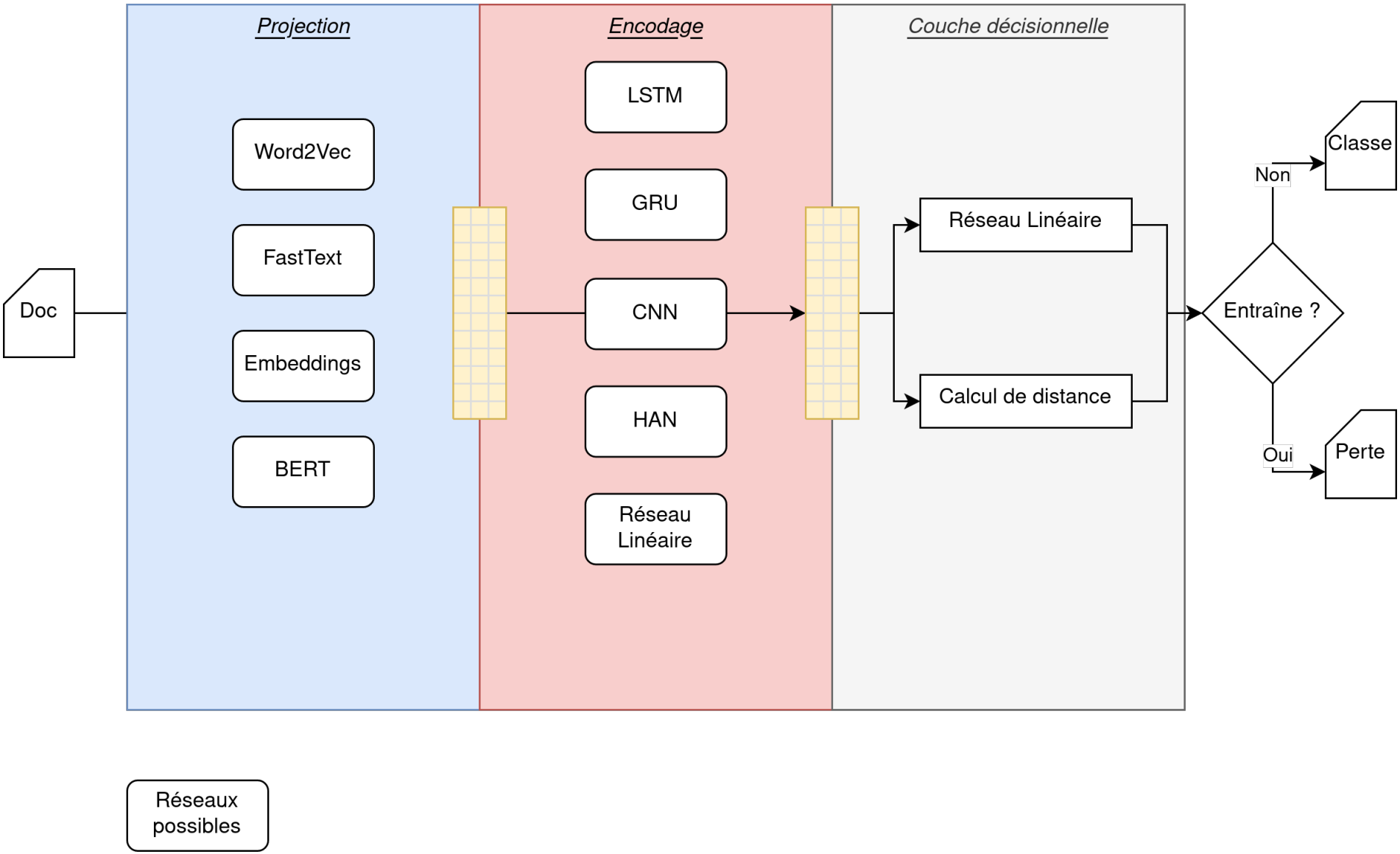
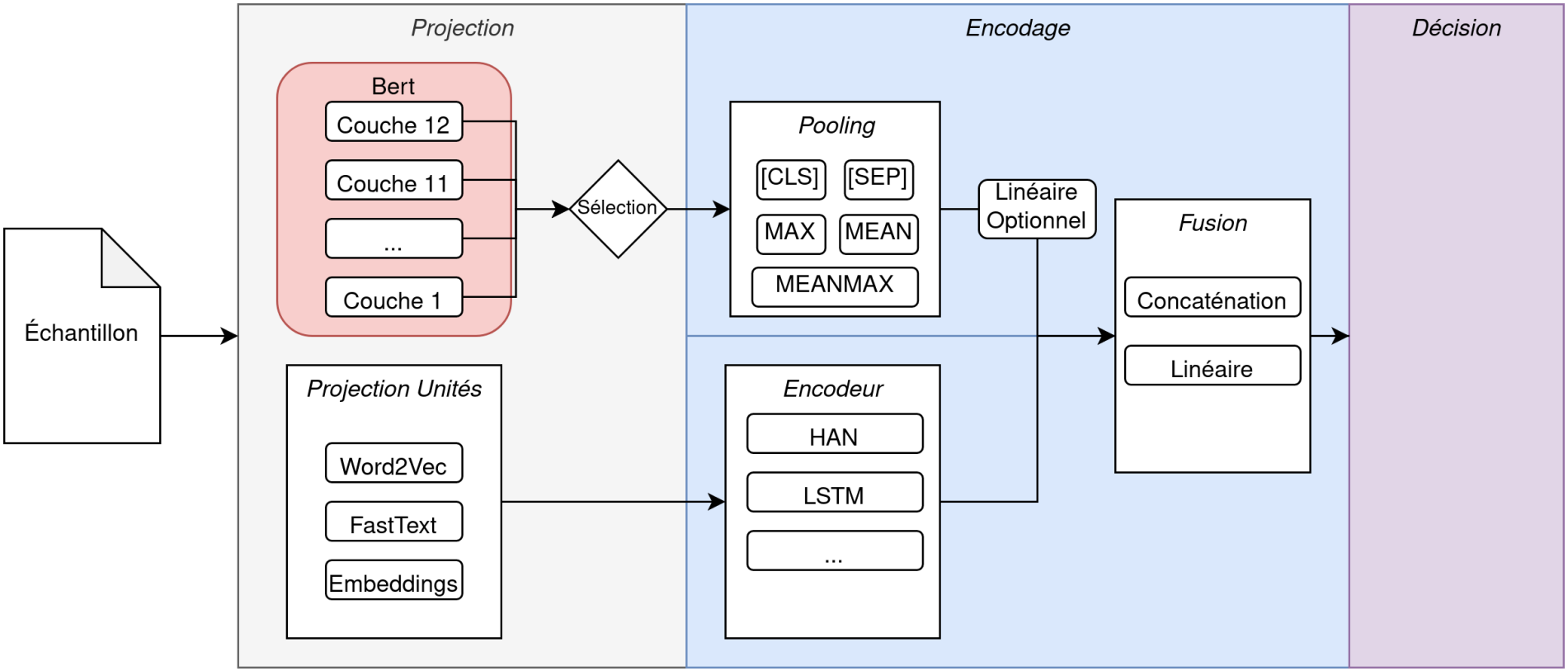


Figure: Accumulation en % du corpus au fil des années. Le meta-corpus est un corpus ouvert et lemmatisé (silver) du 3e siècle BCE au 10e s. CE

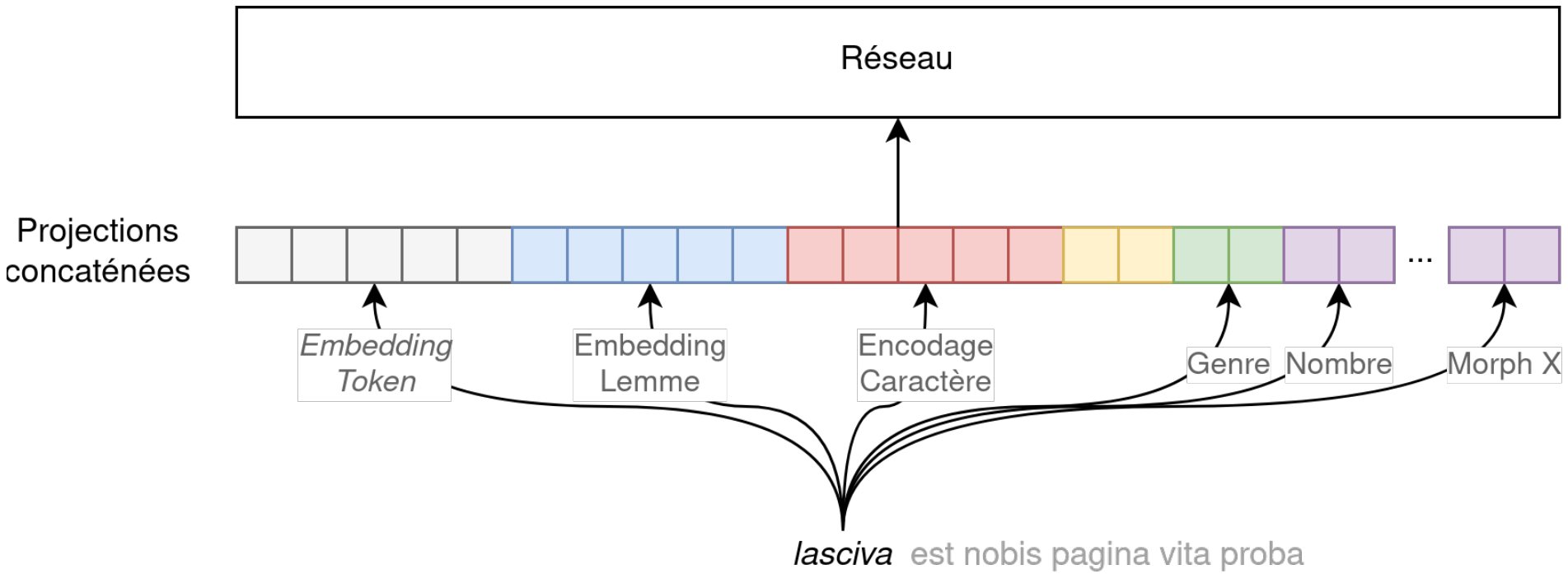
Variation des architectures



Bert


Projection
Encodage
Décision
Bert
Couche 12
Couche 11
...
Couche 1
Sélection
Pooling
[CLS]
[SEP]
MAX
MEAN
MEANMAX
Linéaire Optionnel
Fusion
Concaténation
Linéaire
Échantillon
Projection Unités
Word2Vec
FastText
Embeddings
Encodeur
HAN
LSTM
...

Variation de l'input



Injection de métadonnées

Jihyeok Kim, Reinald Kim Amplayo, Kyungjae Lee, Sua Sung, Minji Seo et Seung-won Hwang, « Categorical metadata representation for customized text classification », *Transactions of the Association for Computational Linguistics*, 7 (2019), p. 201-215.

Injection: LSTM, Attention, Linear

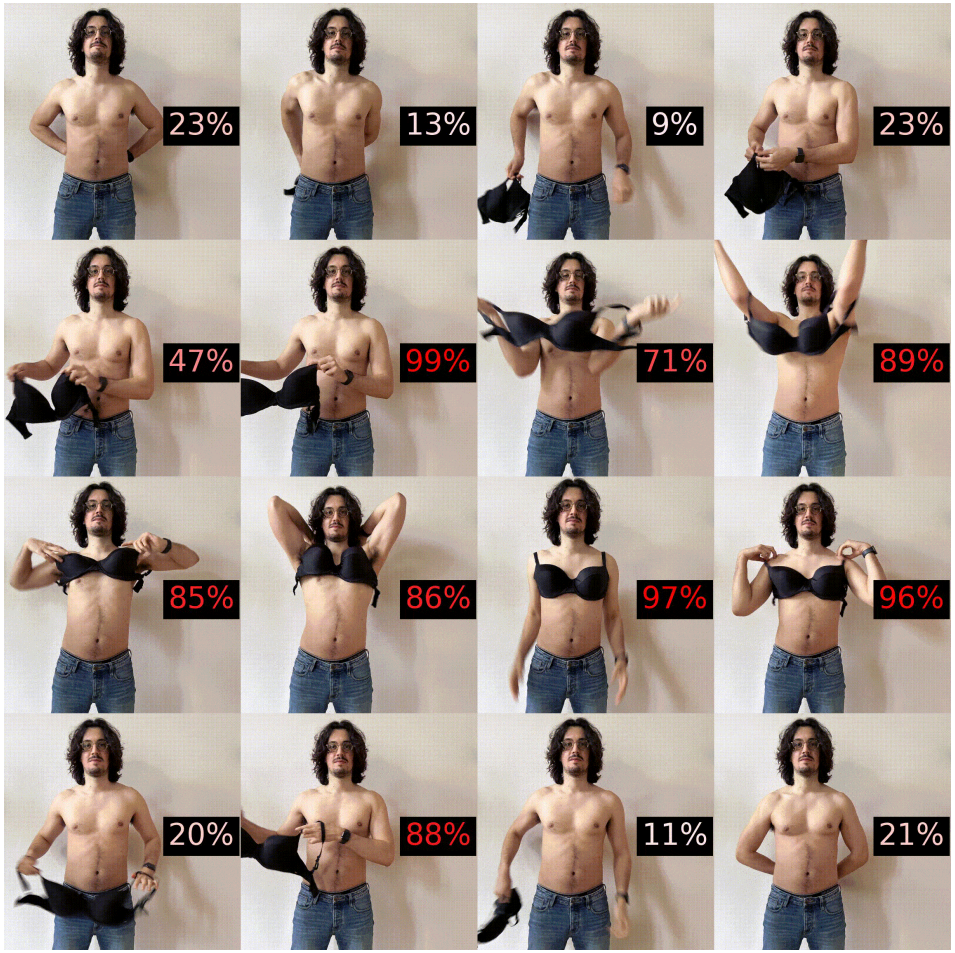
Métadonnées utilisées: Auteur, Structure logique, Forme, Siècle

Globaux

Rang	Score F1 (Positif)	Rappel (Positif)	Précision (Positif)	Couche enrichie	Réseau d'encodage	Métadonnées utilisées
1	86,05%	87,25%	84,88%	Linéaire	HAN	Toutes
2	84,45%	80,08%	89,33%	Linéaire	HAN	Toutes
3	84,15%	88,84%	79,93%	Linéaire	HAN	Toutes
4	84,12%	81,27%	87,18%	Linéaire	HAN	Forme,Siècle,Structure de Citation
5	83,79%	79,28%	88,84%	Linéaire	HAN	Forme,Siècle,Structure de Citation
6	83,30%	78,49%	88,74%	LSTM	LSTM	Toutes
7	83,17%	83,67%	82,68%	Linéaire	HAN	Forme,Siècle,Structure de Citation
8	82,08%	78,49%	86,03%	Linéaire	HAN	Forme,Siècle,Structure de Citation
9	81,82%	78,88%	84,98%	Linéaire	HAN	Toutes
10	81,65%	78,88%	84,62%	LSTM	LSTM	Forme,Siècle,Structure de Citation
11	81,24%	78,49%	84,19%	LSTM	LSTM	Forme,Siècle,Structure de Citation
12	81,08%	77,69%	84,78%	LSTM	LSTM	Forme,Siècle
	...					
23	78,54%	72,91%	85,12%	Aucune		Aucune

Table: Résultats ordonnés par le score F1 de la catégorie positive des meilleurs modèles sur la recherche de meilleure architecture. Les onze premiers modèles utilisent toutes les métadonnées, ou toutes sauf la métadonnée auteur. Les auteurs paramètres n'ont pas d'effet aussi marquant sur le classement.

Robustesse, biais de corpus et sexualité



La photo est-elle osée (*racy*) ?

'There is no standard': investigation finds AI algorithms objectify women's bodies, Gianluca Mauro and Hilke Schellmann, The Guardian, 8 février 2023.

Modèle "inconnu", seul détail "Google's AI" ou "Microsoft's AI"

`https://www.theguardian.com/technology/2023/feb/08/biased-ai-algorithms-racy-women-bodies`

Biais de corpus et littérature latine

Texte	Métadonnées utilisées	En tant que Cicéron	En tant que Martial
De Finibus, Cicéron	Toutes	4,25%	72,60%
Épigrammes, Martial		7,89%	82,05%
De Finibus, Cicéron	Toutes sauf Auteur	4,69%	47,54%
Épigrammes, Martial		7,33%	59,76%
De Finibus, Cicéron	Forme et Siècle	0,91%	36,03%
Épigrammes, Martial		6,41%	48,10%
De Finibus, Cicéron	Aucune	2,91%	2,91%
Épigrammes, Martial		15,29%	15,29%

Table: Nombre de phrases annotées automatiquement comme positives en fonction des modèles avec les meilleurs scores pour chaque type d'usage des métadonnées. L'absence d'usage de métadonnées n'a évidemment aucun impact sur les scores, tandis que l'impact des métadonnées sur le taux de positif est corrélé aux nombres de catégories de métadonnées utilisées.

Conditions

Corpus	Set	train	dev	test
Général	Négatif	19940	2493	2491
Partiel		3970	2493	2491
Métaphores		15701	1745	7478
Non-Métaphores		15701	1745	7478
Général	Positif	2013	252	251
Partiel		420	252	251
Métaphores		413	46	2057
Non-Métaphores		1439	618	459

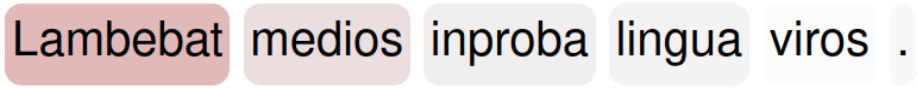
Table: Taille des corpus d'entraînement en fonction de leur composition

Résultats

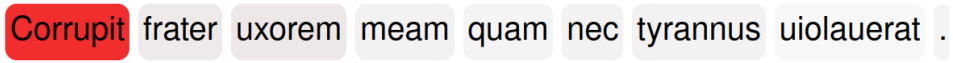
Dataset	Morphologie	Encodeur	Embeddings	Taille encodée	Précision	Rappel	Score F1
Complet	Agglomérée	GRU	Word2Vec	128	**85,07**	71,12	**77,22**
Complet	Agglomérée	HAN	Word2Vec	128	82,38	**71,51**	76,54
Complet	Agglomérée	HAN	Word2Vec	256	84,51	69,52	76,50
Complet	Agglomérée	GRU	Word2Vec	256	84,26	69,72	76,47
Complet	Agglomérée	LSTM	Word2Vec	128	81,80	70,72	74,57
Partiel		HAN	BERT	256	**71,71**	**62,35**	**66,74**
Partiel		GRU	BERT	256	68,70	56,57	62,17
Partiel		MeanMax	BERT	256	**71,72**	50,00	58,92
Partiel	Agglomérée	HAN	Word2Vec	128	66,19	52,59	58,83
Partiel	Agglomérée	GRU	Word2Vec	128	64,66	51,00	56,89
Métaphores	Agglomérée	GRU	Word2Vec	256	94,87	**33,11**	**49,05**
Métaphores	Agglomérée	LSTM	Word2Vec	256	95,36	32,77	48,71
Métaphores	Agglomérée	HAN	Word2Vec	256	95,98	29,39	45,02
Métaphores	Agglomérée	GRU	Word2Vec	128	**96,33**	28,76	44,28
Métaphores	Agglomérée	LSTM	Word2Vec	128	95,86	27,49	42,67
Non-métaphore	Agglomérée	GRU	Word2Vec	256	63,71	61,11	**62,16**
Non-métaphore	Agglomérée	GRU	Word2Vec	128	**64,98**	57,84	60,50
Non-métaphore	Agglomérée	LSTM	Word2Vec	128	61,71	59,59	60,48
Non-métaphore	Agglomérée	HAN	Word2Vec	256	61,89	58,17	59,78
Non-métaphore	Agglomérée	LSTM	Word2Vec	256	57,72	**61,33**	59,38

Table: Top 5 des architectures (Score médian sur 10 runs) hors enrichissement.

Attention et positifs



Martial, *Épigrammes*, II.61.2. "Sa langue perverse léchait le milieu des hommes."



Sénèque l'Ancien, *Controverses*, I.7.4. "Mon frère corrompit ma femme que même le tyran n'avait pas violée."

Différentes répartitions de l'attention en fonction de l'interprétation du modèle. Plus le mot est sur fond foncé, plus il porte d'attention.

Attentions et négatif

Statut Lemme	Faux négatif	Faux positif	Vrai négatif	Vrai positif
!	33,33	0,13	26,97	
.	5,72	0,02	35,63	1,50
;	3,38		22,28	
?	8,18		48,25	0,45
,	0,01	0,00	0,11	0,06

Table: Pourcentage des signes de ponctuation qui se retrouvent en rang 1 de l'attention, sur 20 entraînements différents. Les points d'exclamation se retrouvant en premier rang de l'attention représentent 26,97% des points d'exclamation des échantillons négatifs, qui sont classés en négatifs.

Sommaire

NLP, Philologie et Histoire

On a ici un exemple de chaîne (lemmatisation + classification) qui est proprement de l'ordre du NLP. Mais l'objectif ici est de servir un domaine spécifique (humanités, ici histoire).

Les performances sont à balancer avec l'activabilité des résultats.

Des architectures différentes correspondent à des datasets différents. BERT a pu échouer à cause du modèle original difficile à manipuler et reposant sur des leviers de preprocessing.

Évaluation qualitative en terrain obligatoire pour dépasser le score d'évaluation.

Merci !

Bibliograpie:

`https://github.com/PonteIneptique/these-redaction/tree/master/bibliography`