Retours sur le projet e-CaM

Étiquetage lexico-grammatical du castillan médiéval

Olivier Brisville-Fertin, CIHAM-UMR 5648 & ENS de Lyon
Matthias Gille Levenson, CIHAM-UMR 5648 & UVSQ

Symposium Agorantic, 1er décembre 2025

Introduction : l’annotation lexico-grammaticale

L'annotation grammaticale est une tâche linguistique double. Elle vise à:

- réduire chaque “mot” à sa forme standard, en général son entrée de dictionnaire: le **lemme**
- lui attribuer une ou plusieurs étiquettes décrivant l’information grammaticale et morphologique de la forme (nature grammaticale, genre, nombre, déclinaison, etc): c’est le **PoS** (*Part of Speech*) et la **morphologie**.

Utilités de l’annotation lexico–grammaticale:

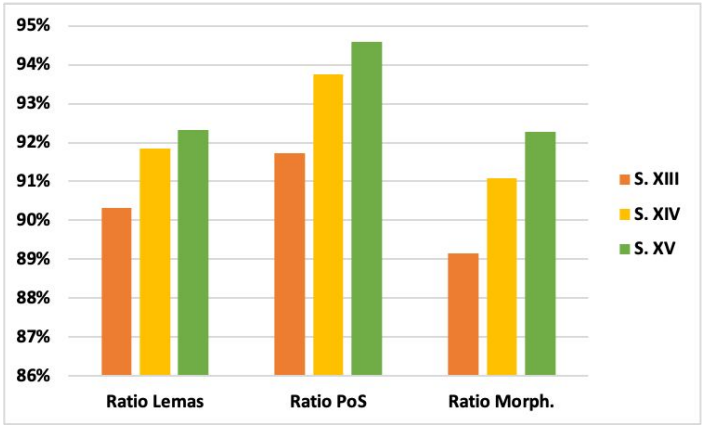
- en philologie (collation automatisée)
- pour faciliter l'étude du texte médiéval (études de thèmes)
- en linguistique de corpus
- pour pousser les études vers la syntaxe (production de *treebanks*)

Introduction : état de l’art

- Plusieurs corpus en ligne non lemmatisés (e.g. CORDE) et lemmatisés (CODEA+, OSTA, CDHLE)
- Un seul outil d’annotation du castillan libre et documenté : adaptation de *Freeling* [Sánchez Marco 2012]. Utilisation de listes annotées et de Modèles cachés de Markov pour l'annotation lexicale et morphologique.
- Apparition de nouveaux outils d'annotation lexicale comme *Pie* [Manjavacas *et al.* 2019] fondés sur les réseaux de neurones profonds permettant d'améliorer l'annotation en contexte

Introduction: évaluation de Freeling

M. Gille Levenson, O. Brisville-Fertin, M.ª Díez Yáñez, S. Gabay, « Construcción de un corpus de evaluación de la anotación léxico-gramatical del castellano medieval (siglos XIII-XV) », communication *Humanidades Digitales Hispánicas* 2021 (en ligne) ; dépôt HAL.



- **Taille de corpus** de 35 k tokens, 166 œuvres (13e-15e siècles) corrigé manuellement pour évaluation.
- **Résultats** : lemmes, 91,61 % ; PoS, 92,97 % ; Morph, 90,61 %.
  - 2 269 erreurs de lemmes : 36,6 % autres lemmes; 62,8 % lemmes faux ⇒ **variation (ortho)graphique** ;
  - erreurs de nature ou PoS (7,03 %) : surtout pour les noms (9,5 % faux) et les adjectifs (7 % faux) ⇒ liées au contexte ;
  - erreurs morphologiques (9,39 %) : surtout le verbes (20,95 % faux) ⇒ **homographies = ambiguïtés hors contexte**.
- Problèmes du jeu d'étiquettes (EAGLES) : **limites** et analyses problèmatiques.

Introduction : le projet e-CaM

Objectifs généraux : améliorer l'annotation lexico-grammatical

- Construction d'un corpus représentatif du castillan médiéval (Phase 1)
- Annotation de ce corpus et réflexion du passage à *Universal Dependencies* (Phase 1 + Phase 2)
- Documentation des normes d'annotations dans un manuel collaboratif (Phase 1)
- Entraînement d'un modèle performant et dépassant l'existant (Phase 2)

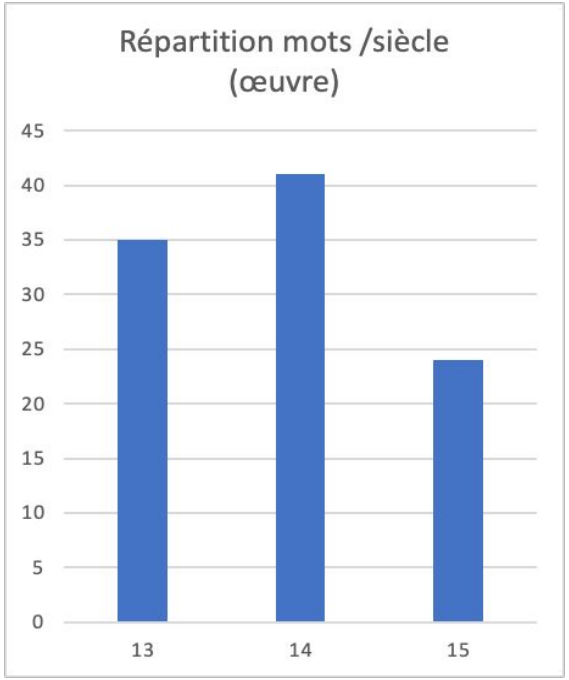
Résultats annoncés et atteints pour la phase 1 proposée à Agorantic :

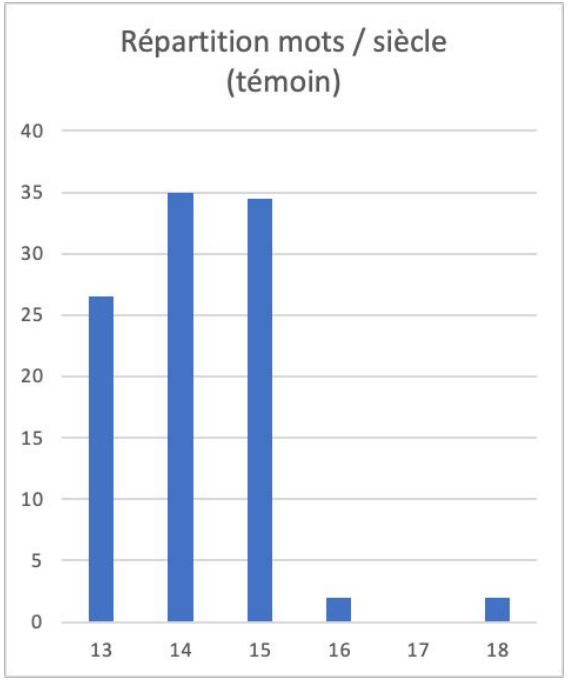
- Protocole d'étiquetage documenté dans un manuel ;
- Correction d'un 1er corpus de 50 k mots.

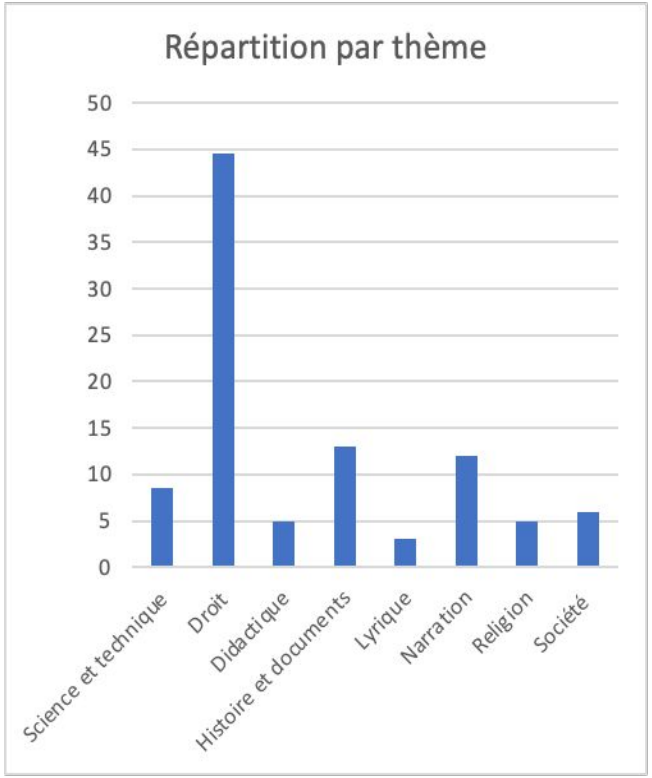
Le projet : constitution d’un corpus

- Un corpus global de 1,3 million de tokens
- Récupérés par *scrapping* à partir de la base de données linguistique CORDE (*CORpus Diacrónico del Español*, Real Academia Española)
- Le corpus n'a pas été équilibré manuellement, car il nous semblait relativement satisfaisant (stats. *infra*).
- Le corpus sera divisé en **10 sous-corpus** et annoté par campagnes successives.
- Le corpus est disponible sous divers formats, **chaque fragment est référencé**.
- En moyenne : **100** tokens / exemple et **12** exemples / oeuvre.

Le projet : statistiques du corpus







- Équilibre relatif entre siècles
- Équilibre entre textes littéraires et documents de la pratique (notariés).

Le projet : des normes de transcription variées

- Les exemples sont tirés d'éditions scientifiques couvrant tout le XX^{e} siècle (prédominance *post* 1950)
- Un grand nombre de normes de transcriptions est ainsi représenté dans le corpus

Le projet : le choix du référentiel lexical

- Lemmatisation de Freeling parfois questionnable :(*conquistar*/*conquerir*)
- Choix de référentiel : dictionnaire de l'Académie royale de la langue (RAE)
- Protocole d'ajout de lemmes inconnus :
  - 23^{e} éd. du dictionnaire de l'Académie
  - versions antérieures ;
  - références du *Tesoro Lexicográfico de la Lengua Española* ;
  - *Diccionario del español medieval* ;
  - Création *ad hoc* d'une entrée dans le référentiel

Le projet : le choix du jeu d'étiquettes EAGLES ou *UD* ?

EAGLES (années 1990) :

- Étiquettes abrégées fixes de lettres et chiffres : N(ature)A(ttribut)V(aleur), etc.
- 13 catégories de nature pouvant avoir jusqu'à 7 valeurs possibles
  - *Niñas* : "niño" NCFP000 ⇔ N(om)C(ommun)F(éminin)P(luriel)000
  - *Cantan* : "cantar" VMIP3P0 ⇔ V(erbe)M(principal)I(ndicatif)P(résent)3(e pers.)P(luriel)0.
- Choix critiquables ou limités :
  - *Usted*: "tú" PP2CS0P ⇔ P(ron.)P(ers.)2(pers.)C(ommune)S(ing.)0P(olitesse)
  - *Cómo* : "cómo" PT00000 ⇔ P(ron.)T(interrogatif)00000

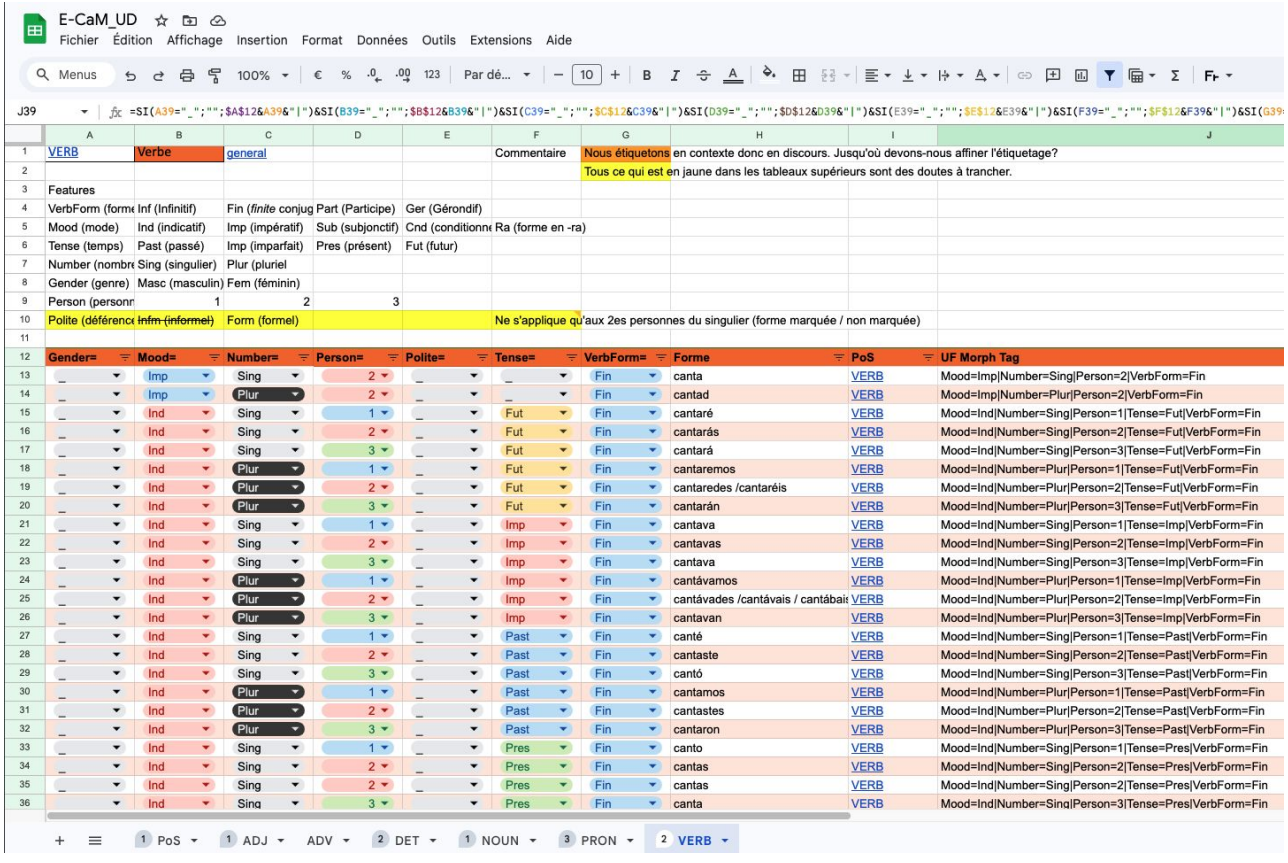
Universal Dependencies (années 2010) :

- Étiquettes morphologiques par concaténation d'attributs : NATURE Att1:val|Att2:val|etc.
  - *Cantan* : "cantar" VERB Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin.
- Arbres de dépendances des relations syntaxiques (potentiel futur)

Passage à UD avec sélection réfléchie des catégories et traits conservés pour e-CaM.

Le projet : le choix de *Universal Dependencies*

Établissement d’ étiquettes possibles pour les natures aux attributs complexes



E-CaM_UD

Fichier Édition Affichage Insertion Format Données Outils Extensions Aide

	A	B	C	D	E	F	G	H	I	J
1	VERB	Verbe	general			Commentaire	Nous étiquetons	en contexte donc en discours. Jusqu’où devons-nous affiner l’étiquetage?		
2							Tous ce qui est	en jaune dans les tableaux supérieurs sont des doutes à trancher.		
3	Features									
4	VerbForm (forme	Inf (Infinitif)	Fin (*finite* conjug	Part (Participe)	Ger (Gérondif)					
5	Mood (mode)	Ind (indicatif)	Imp (impératif)	Sub (subjonctif)	Cnd (conditionne	Ra (forme en -ra)				
6	Tense (temps)	Past (passé)	Imp (imparfait)	Pres (présent)	Fut (futur)					
7	Number (nombre	Sing (singulier)	Plur (pluriel							
8	Gender (genre)	Masc (masculin)	Fem (féminin)							
9	Person (personn	1	2	3						
10	Polite (déférenc	~~Infm (informel)~~	Form (formel)			Ne s’applique qu’	aux 2es personnes du singulier (forme marquée / non marquée)			
11										

12	Gender=	Mood=	Number=	Person=	Polite=	Tense=	VerbForm=	Forme	PoS	UF Morph Tag
13	_	Imp	Sing	2	_	_	Fin	canta	VERB	Mood=Imp|Number=Sing|Person=2|VerbForm=Fin
14	_	Imp	Plur	2	_	_	Fin	cantad	VERB	Mood=Imp|Number=Plur|Person=2|VerbForm=Fin
15	_	Ind	Sing	1	_	Fut	Fin	cantaré	VERB	Mood=Ind|Number=Sing|Person=1|Tense=Fut|VerbForm=Fin
16	_	Ind	Sing	2	_	Fut	Fin	cantarás	VERB	Mood=Ind|Number=Sing|Person=2|Tense=Fut|VerbForm=Fin
17	_	Ind	Sing	3	_	Fut	Fin	cantará	VERB	Mood=Ind|Number=Sing|Person=3|Tense=Fut|VerbForm=Fin
18	_	Ind	Plur	1	_	Fut	Fin	cantaremos	VERB	Mood=Ind|Number=Plur|Person=1|Tense=Fut|VerbForm=Fin
19	_	Ind	Plur	2	_	Fut	Fin	cantaredes /cantaréis	VERB	Mood=Ind|Number=Plur|Person=2|Tense=Fut|VerbForm=Fin
20	_	Ind	Plur	3	_	Fut	Fin	cantarán	VERB	Mood=Ind|Number=Plur|Person=3|Tense=Fut|VerbForm=Fin
21	_	Ind	Sing	1	_	Imp	Fin	cantava	VERB	Mood=Ind|Number=Sing|Person=1|Tense=Imp|VerbForm=Fin
22	_	Ind	Sing	2	_	Imp	Fin	cantavas	VERB	Mood=Ind|Number=Sing|Person=2|Tense=Imp|VerbForm=Fin
23	_	Ind	Sing	3	_	Imp	Fin	cantava	VERB	Mood=Ind|Number=Sing|Person=3|Tense=Imp|VerbForm=Fin
24	_	Ind	Plur	1	_	Imp	Fin	cantávamos	VERB	Mood=Ind|Number=Plur|Person=1|Tense=Imp|VerbForm=Fin
25	_	Ind	Plur	2	_	Imp	Fin	cantávades /cantávais / cantábais	VERB	Mood=Ind|Number=Plur|Person=2|Tense=Imp|VerbForm=Fin
26	_	Ind	Plur	3	_	Imp	Fin	cantavan	VERB	Mood=Ind|Number=Plur|Person=3|Tense=Imp|VerbForm=Fin
27	_	Ind	Sing	1	_	Past	Fin	canté	VERB	Mood=Ind|Number=Sing|Person=1|Tense=Past|VerbForm=Fin
28	_	Ind	Sing	2	_	Past	Fin	cantaste	VERB	Mood=Ind|Number=Sing|Person=2|Tense=Past|VerbForm=Fin
29	_	Ind	Sing	3	_	Past	Fin	cantó	VERB	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin
30	_	Ind	Plur	1	_	Past	Fin	cantamos	VERB	Mood=Ind|Number=Plur|Person=1|Tense=Past|VerbForm=Fin
31	_	Ind	Plur	2	_	Past	Fin	cantastes	VERB	Mood=Ind|Number=Plur|Person=2|Tense=Past|VerbForm=Fin
32	_	Ind	Plur	3	_	Past	Fin	cantaron	VERB	Mood=Ind|Number=Plur|Person=3|Tense=Past|VerbForm=Fin
33	_	Ind	Sing	1	_	Pres	Fin	canto	VERB	Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin
34	_	Ind	Sing	2	_	Pres	Fin	cantas	VERB	Mood=Ind|Number=Sing|Person=2|Tense=Pres|VerbForm=Fin
35	_	Ind	Sing	2	_	Pres	Fin	cantas	VERB	Mood=Ind|Number=Sing|Person=2|Tense=Pres|VerbForm=Fin
36	_	Ind	Sing	3	_	Pres	Fin	canta	VERB	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin

PoS ADJ ADV DET NOUN PRON VERB

Le projet : annotation préalable du corpus

- Corpus annoté automatiquement avec Freeling.
- Les étiquettes grammaticales et morphologiques sont ensuite converties en UD via un script *ad hoc*, à partir d'une table de correspondance récupérée sur le projet ANCORA [Martínez Alonso *et al.* 2016]
- Les sous-corpus sont produits et chargés dans *Pyrrha* [Clérice *et al.* 2022].

Le projet : première campagne d’annotation
Résultats : un premier corpus corrigé

- Deux vacataires (niveau M2, spécialisés en linguistique diachronique) ont été recrutés pour la première campagne (**100 h**).
- ≈ **50.000** tokens corrigés : premier corpus embryonnaire
- Le protocole d’annotation et choix du jeu d’étiquette ont pu être rediscutés en atelier

Résultats : le manuel d’annotation

Rédaction d’un manuel documentant le protocole d’annotation (après les retours de la campagne de correction)

Quelques exemples de choix

- Lexique :
  - Double lemmatisation : lemme conservateur + lemme modernisé/régularisé
    - ***Conquiso*** *Calaforra* : “conquerir” | “conquistar”
  - Désambiguïsation du lemme modernisé
    - ***e y*** *encontró* : “y” | “y2” ; “y” | “y3” [Gabay *et al.* 2025]
- Morphologie :
  - Participes : étiquetés comme des verbes
  - Forme en *-ra* (subj. impf.) : temps *ad hoc* sans mode Ra
    - ***Matáran****le los hermanos* : VERB Number=Plur|Person=3|Tense=Ra|VerbForm=Fin
  - Abandon du trait politesse : ambiguïté en raison des contextes réduits

Résultats : limites et questions

- Correction en lot *versus* annotations en campagnes
- Volume critique de données afin d'améliorer les scores ?

Résultats à venir : vers l’entraînement d'un modèle d’annotation performant

- On utilisera l’outil *Pie* afin d'entraîner un nouveau modèle
- Développement envisagé d'une solution mixte: *Pie* U dictionnaire de formes annotées (si pas d'ambiguïté) disponible via *Freeling*.
- Un *datapaper* est prévu à l’issue du projet.

Conclusion

- Première phase exploratoire pour l’amélioration de l’étiquetage lexico-grammaticale
- Base expérimentale pour une poursuite financée par Biblissima+ : e-CaM^2

Merci à Agorantic pour sa confiance

Merci de votre attention !

Bibliographie

- Clérice, Thibault, Vincent Jolivet, et Julien Pilla. “Building Infrastructure for Annotating Medieval, Classical and Pre-Orthographic Languages: The Pyrrha Ecosystem.” *DH2022*, 2022.
- Simon Gabay, Lucence Ing, Ariane Pinche et Sonia Solfrini, *Guide pour le traitement numérique des textes en français*, 2025 (à paraître)
- Manjavacas, Enrique, Ákos Kádár, et Mike Kestemont. “Improving Lemmatization of Non-Standard Languages with Joint Learning.” *Proceedings of the 2019 Conference of the North*, Association for Computational Linguistics, 2019, 1493–503. https://doi.org/10.18653/v1/N19-1153.
- Martínez Alonso, Héctor et Daniel Zeman. “Universal Dependencies for the AnCora Treebanks.” *Procesamiento Del Lenguaje Natural*, no. 57 (2016). https://inria.hal.science/hal-01426751/.
- Sánchez Marco, Cristina. “Tracing the Development of Spanish Participial Constructions: An Empirical Study of Semantic Change.” Thèse de Doct., Universitat Pompeu Fabra, 2012. https://www.tdx.cat/bitstream/handle/10803/97044/tcsm.pdf?sequence=1.