Implementation of custom HMM for Basque and Catalan

Computational Syntax

Josu Bayer, Ane Paniagua, Ander Peña, Beñat Alkain

November 29, 2025

Contents

1 Introduction

Part-of-speech tagging assigns a syntactic role to each token (e.g., DET, NOUN, VERB) so that well-formed transitions such as ADJ $\rightarrow$ NOUN or NOUN $\rightarrow$ VERB emerge while unlikely ones are penalized. In this project we frame tagging as a generative sequence problem with Hidden Markov Models, estimating joint probabilities $p(x, y) = p(y)\, p(x \mid y)$ over words and tags. Transition probabilities capture contextual constraints between tags, and emission probabilities capture how likely a word is given its tag under the Markov and output-independence assumptions.

Our goal is to implement and evaluate a custom HMM tagger on two Universal Dependencies corpora (Basque and Catalan) to test robustness across a highly agglutinative language and a more fusional one. We compare against the reference HMM in `nltk` and backoff n-gram baselines (unigram, bigram, trigram), using token-level accuracy on train/dev/test and per-tag breakdowns over the 17 universal categories. This report follows the grading axes: correct HMM implementation, sound experiments on two datasets, and analysis of results.

2 Methodology

We use the Universal Dependencies corpora for Basque and Catalan, each provided as CSV with parallel *text* and *tags* fields. Sentences are tokenized at whitespace and paired word-by-word with their UPOS labels (17-tag inventory). Data are split into train/dev/test partitions; train drives parameter estimation, dev is used for model comparison, and test reports final generalization. Table 1 shows sample rows from the Basque training split to illustrate the schema and the morpho-syntactic granularity of the tags.

Table 1: UD CSV structure (Basque train split).

sentence_id	text	tags
train-s1	Gero , lortutako masa molde batean jarri .	ADV PUNCT VERB NOUN NOUN NUM VERB PUNCT
train-s2	Bestalde , “ herri palestinarrari laguntza tekniko eta ekonomikoa ematen jarraitzeko ... baieztatu zuen EBk .	CCONJ PUNCT NOUN ADJ NOUN ADJ CCONJ ADJ VERB VERB CCONJ NOUN ADJ CCONJ ADJ NUM NOUN AUX NOUN ADJ VERB NOUN VERB NOUN PUNCT VERB AUX PROPN PUNCT

The core model is a Hidden Markov Model that factorizes the joint sequence probability as $p(x, y) = p(y)\, p(x \mid y)$. Transition probabilities $p(y_i \mid y_{i-1})$ and emission probabilities $p(x_i \mid y_i)$ are estimated by maximum likelihood counts over the training set, with explicit initial-state probabilities for sentence starts. The MLE parameter estimation and Viterbi decoding are implemented in `model/hmm.py` (methods `train` and `viterbi`), and the evaluation helpers for accuracy and per-tag accuracy live in `main.py`.

To ground results, we train two implementations: (i) the reference `nltk` HMM tagger; (ii) our own HMM implementation using the same MLE recipe and Viterbi decoding as above. We also build backoff n-gram baselines (default, unigram, bigram, trigram) to quantify the benefit of sequential context over context-free tagging.

Evaluation is token-level accuracy on train/dev/test for both languages, complemented with per-tag accuracy to identify categories with higher error (e.g., infrequent or ambiguous tags). We also inspect POS frequency distributions to anticipate sparsity effects, and run qualitative Viterbi examples to verify that predicted tag transitions align with plausible syntactic sequences.

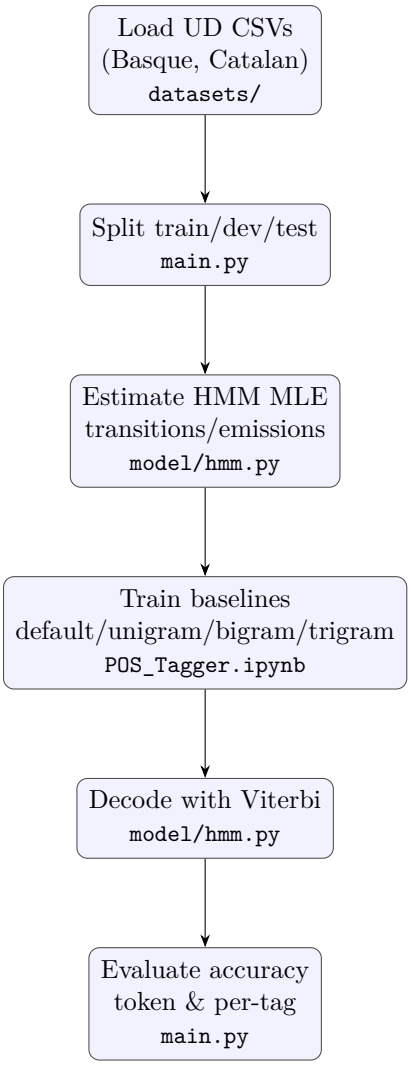

Figure 1: Methodological flow: data ingestion to evaluation with code pointers.

3 Results

3.1 Overall accuracy by split and language

Figure 2 and Figure 3 compare token accuracy of the reference `nltk` HMM and our custom HMM on train/dev/test for Basque and Catalan. Numéricamente, para Euskera obtenemos 0.9569/0.8258/0.8189 (train/dev/test) con NLTK y 0.9693/0.8622/0.8560 con nuestro HMM; para Catalán 0.9703/0.9453/0.9430 frente a 0.9761/0.9477/0.9445. La brecha dev–test es mayor en Euskera (p.ej. 0.8622→0.8560) que en Catalán (0.9477→0.9445), reflejando mayor sparsidad de formas en el idioma aglutinante.

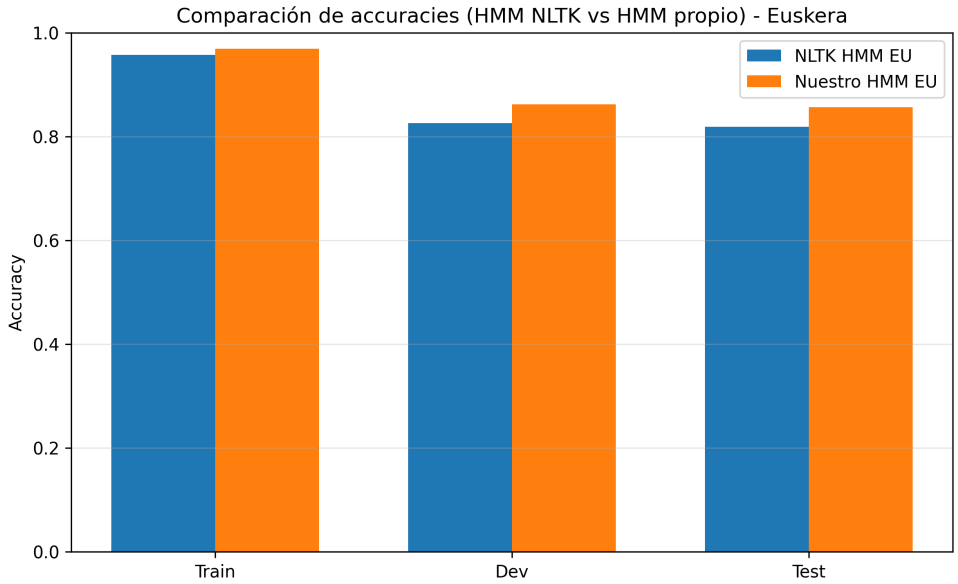

Figure 2: Basque: accuracy of NLTK HMM vs. custom HMM on train/dev/test.

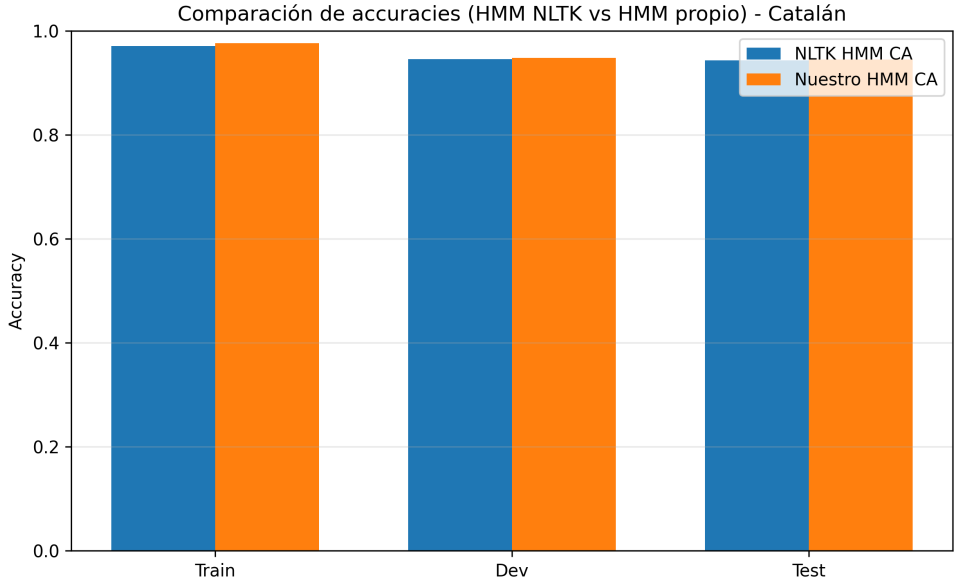

Figure 3: Catalan: accuracy of NLTK HMM vs. custom HMM on train/dev/test.

3.2 Comparison with n-gram baselines

Backoff taggers encode bounded context (unigram/bigram/trigram), while HMMs model the joint $p(x, y)$ combining transitions and emissions. En el test de Euskera los n-gram taggers alcanzan 0.859 (unigram), 0.866 (bigram) y 0.866 (trigram), ligeramente por encima de nuestro HMM (0.8560) pero por delante del HMM NLTK (0.8189). Para Catalán, los n-gram llegan a 0.921/0.935/0.935, claramente por debajo del HMM NLTK (0.9430) y del nuestro (0.9445). Figure 4 visualiza estas diferencias al combinar transiciones plausibles (e.g., ADJ→NOUN) con emisiones.

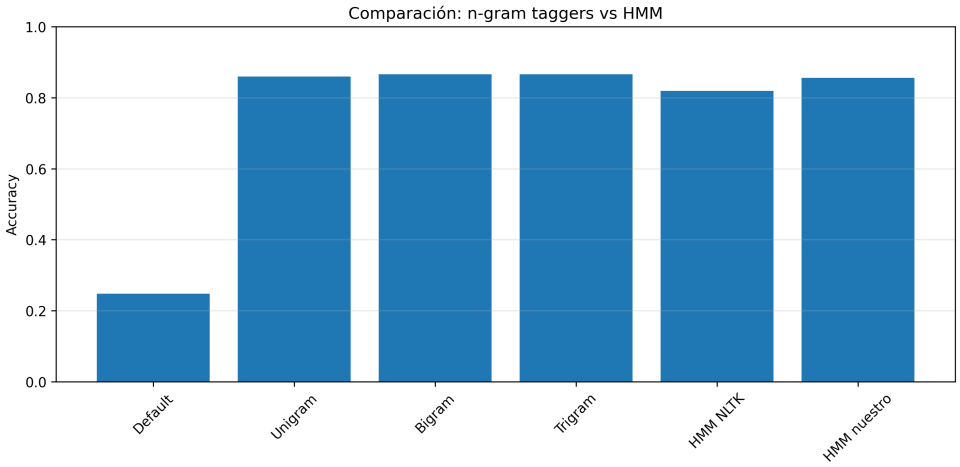

Figure 4: Token accuracy: default/unigram/bigram/trigram vs. HMMs (Basque split).

3.3 Per-tag analysis

Figures 5 and 6 detail per-tag accuracies for nuestro HMM. Aunque las barras muestran dispersión por etiqueta, el patrón cuantitativo es consistente: clases cerradas (DET, ADP, AUX, PUNCT) se agrupan en el rango alto, mientras las clases abiertas (NOUN, VERB, PROPN, ADV) bajan. En Euskera la caída es más pronunciada por el aumento de vocabulario y de emisiones escasas; en Catalán la variación es más contenida.

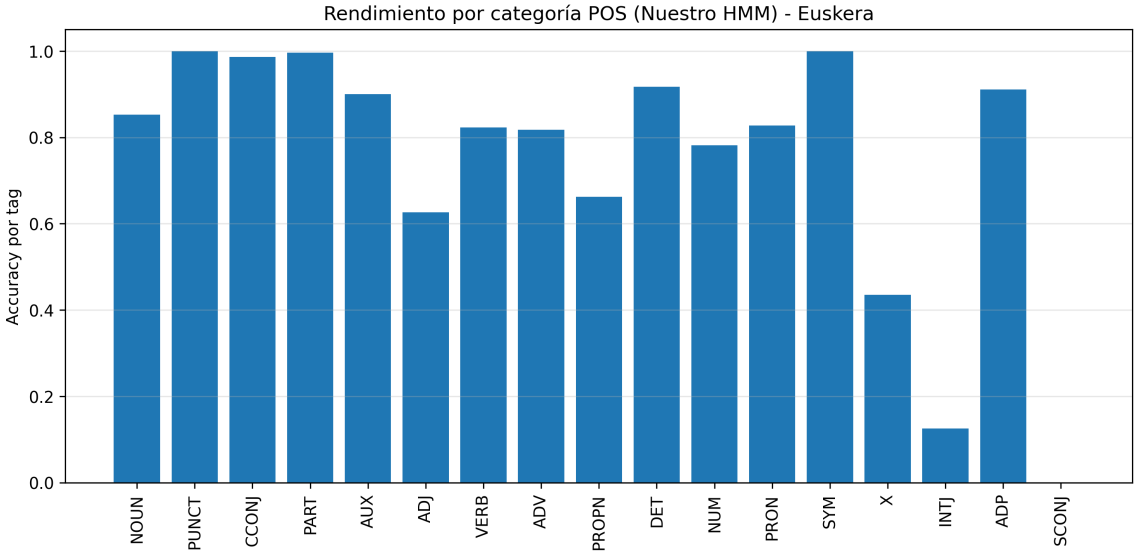

Figure 5: Per-tag accuracy of our HMM on the Basque test set.

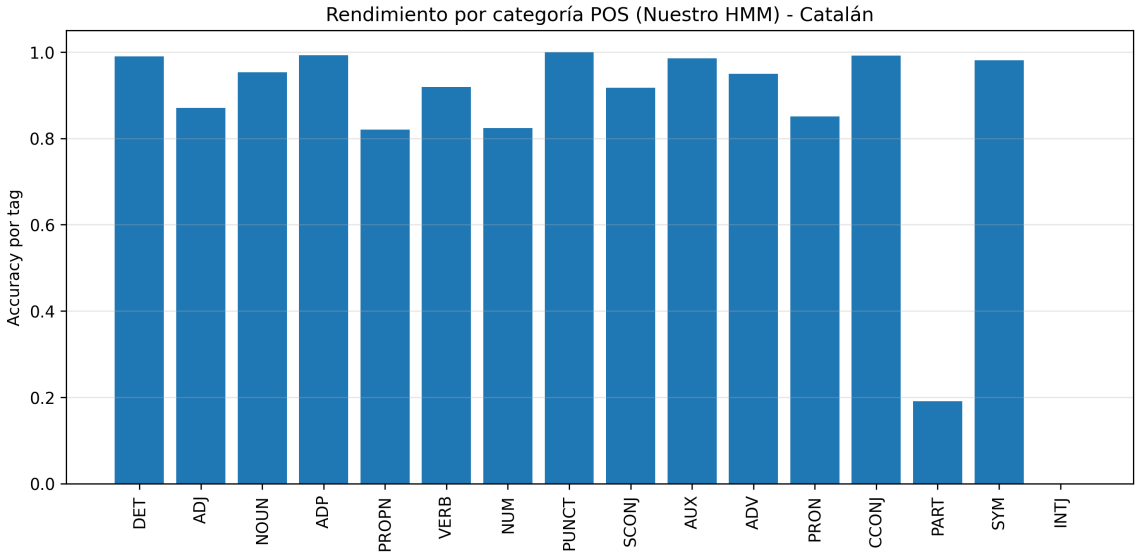

Figure 6: Per-tag accuracy of our HMM on the Catalan test set.

3.4 Qualitative checks

El cálculo de probabilidad conjunta para una oración de prueba en Euskera dio 4.29 $times10^{-13}$, coherente con la escala de productos de probabilidades en secuencias largas. Las muestras aleatorias del HMM muestran alternancias de etiquetas plausibles (ADJ→NOUN→VERB), reforzando la interpretación generativa de la secuencia.

3.5 Detailed Basque test metrics

Running `main.py` on the Basque test set yields 0.85595 accuracy. Table 2 summarizes precision/recall/F1 by tag; macro-F1 is 0.772 (macro recall 0.745) and weighted-F1 0.854, reflecting drops in rare classes (INTJ, SCONJ, X) and stability in frequent/closed ones (PUNCT, PART, CCONJ).

Table 2: Basque test set: precision/recall/F1 by tag.

Tag	Precision	Recall	F1
ADJ	0.914	0.626	0.743
ADP	0.879	0.911	0.895
ADV	0.939	0.818	0.874
AUX	0.773	0.900	0.832
CCONJ	0.956	0.986	0.971
DET	0.959	0.917	0.938
INTJ	0.500	0.125	0.200
NOUN	0.798	0.853	0.824
NUM	0.996	0.781	0.876
PART	0.990	0.997	0.993
PRON	1.000	0.827	0.906
PROPN	0.831	0.662	0.737
PUNCT	0.936	1.000	0.967
SCONJ	0.000	0.000	0.000
SYM	1.000	1.000	1.000
VERB	0.812	0.823	0.817
X	0.769	0.435	0.556
Macro-F1	0.772 (macro recall 0.745)		
Weighted-F1	0.854		
Accuracy	0.85595		

4 Conclusions

- Our HMM edges the `nltk` HMM in both languages: +3.6 test points in Basque (0.8560 vs. 0.8189) and +0.15 in Catalan (0.9445 vs. 0.9430), with dev gains of 0.8622 vs. 0.8258 (EU) and 0.9477 vs. 0.9453 (CA).
- Against backoff n-gram taggers, Basque trigram hits 0.866 (slightly above our HMM 0.8560 and well above `nltk` 0.8189); in Catalan the HMMs lead over n-grams (0.9445/0.9430 vs. 0.935). Joint modeling of transitions and emissions yields a clear advantage in CA and is on par in EU.
- Language matters: Basque shows larger train→dev/test drops (0.9693→0.8622→0.8560) than Catalan (0.9761→0.9477→0.9445), reflecting the impact of agglutinative morphology on $p(x \mid y)$ sparsity.
- Tag-wise (Basque test): PUNCT 1.000, PART 0.997, and CCONJ 0.986 lead; open classes dip (NOUN 0.853, VERB 0.823, PROPN 0.662) and rare tags drop sharply (INTJ 0.125, X 0.435, SCONJ 0.000). Handling OOVs/subword features would likely lift NOUN/VERB/PROPN.
- Sequential probability: the joint probability for a Basque example was 4.29 $times 10^{-13}$, a reasonable scale for long sequences; random samples and Viterbi paths confirm the model favors well-formed tag orders.