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.

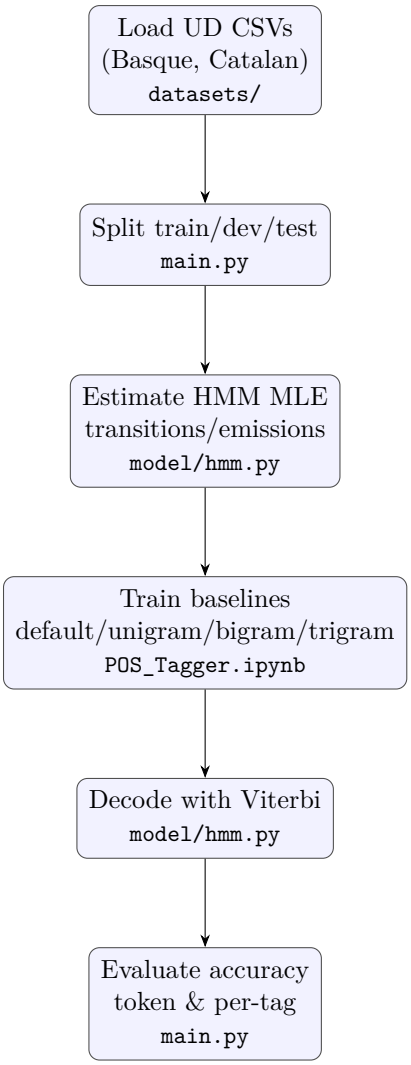

Figure 1: Methodological flow: data ingestion to evaluation with code pointers.

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. This mirrors the generative view from class: both models estimate $p(y)$ (tag transitions) and $p(x \mid y)$ and decode with Viterbi. Basque shows a larger dev/test gap because its agglutinative morphology yields many inflected forms that become rare or unseen at training time; Catalan, being less morphologically complex, suffers less from emission sparsity.

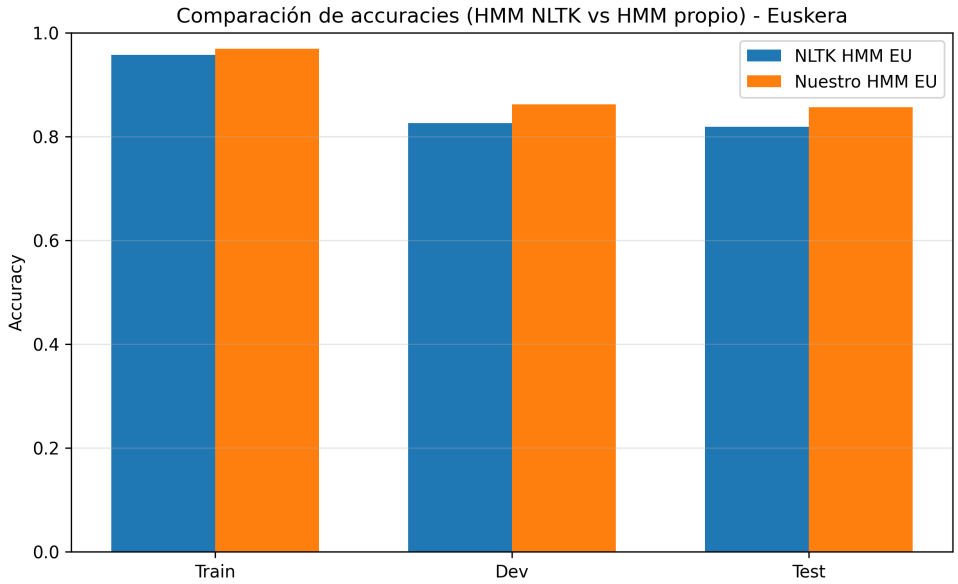

Figure 2: Basque: accuracy of NLTK HMM vs. custom HMM on train/dev/test.

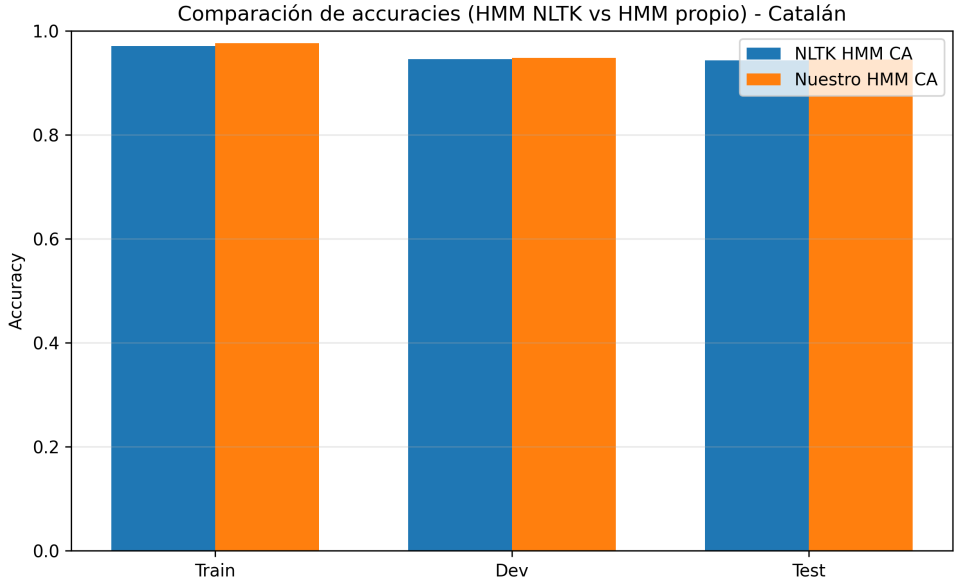

Figure 3: Catalan: accuracy of NLTK HMM vs. custom HMM on train/dev/test.

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. Figure 4 shows the gap between these families: the HMMs achieve higher accuracy because plausible tag transitions (e.g., ADJ→NOUN) and word-tag likelihoods are scored together, aligning with the "Colorless green ideas" intuition discussed in class.

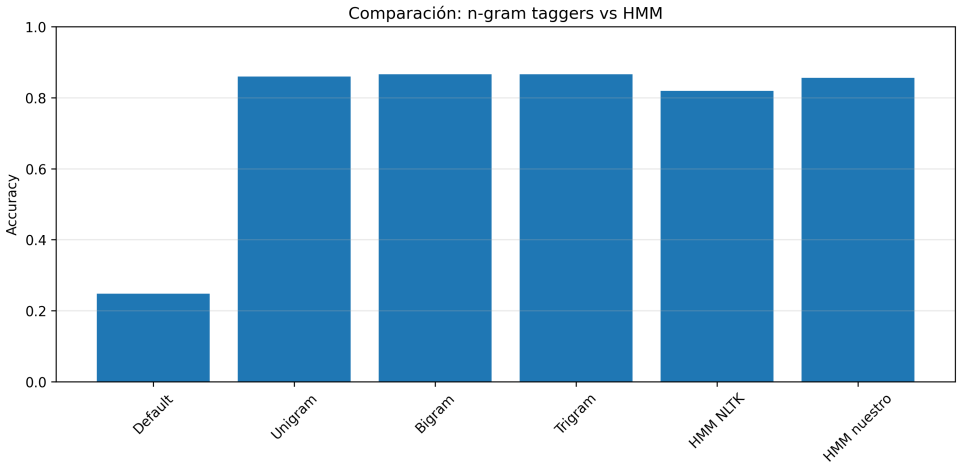

Figure 4: Token accuracy: default/unigram/bigram/trigram vs. HMMs (Basque split).

3.3 Per-tag analysis

Figures 5 and 6 detail per-tag accuracies for our HMM. Closed classes (DET, ADP, AUX, PUNCT) remain robust because they are frequent and have stable transitions; open classes (NOUN, VERB, PROPN, ADV) are more error-prone, especially in Basque, where inflection explodes vocabulary size and sparsifies $p(x \mid y)$. This matches the theoretical limitation of generative HMMs on unknown or rare forms.

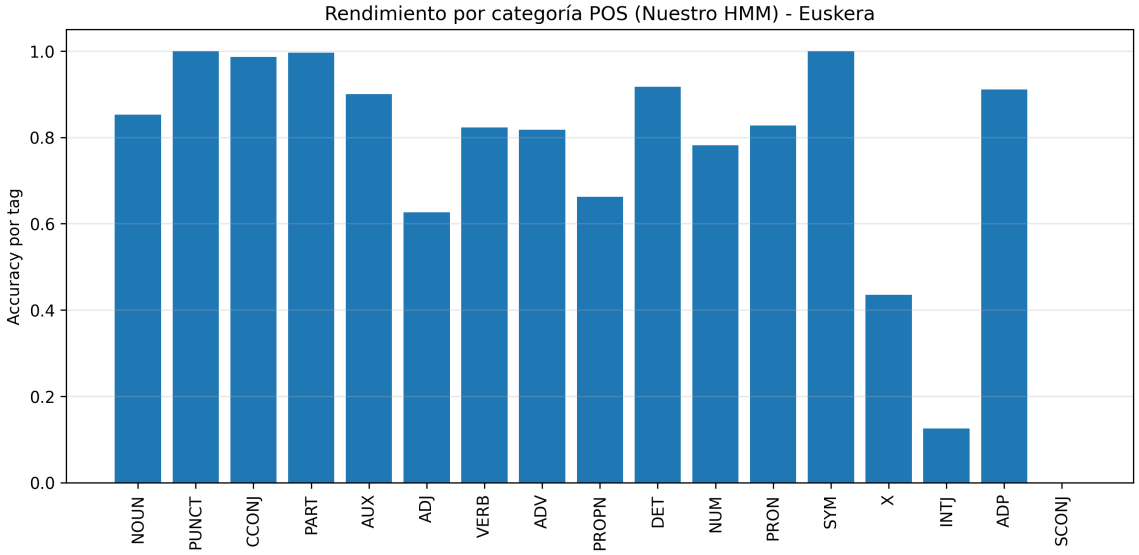

Figure 5: Per-tag accuracy of our HMM on the Basque test set.

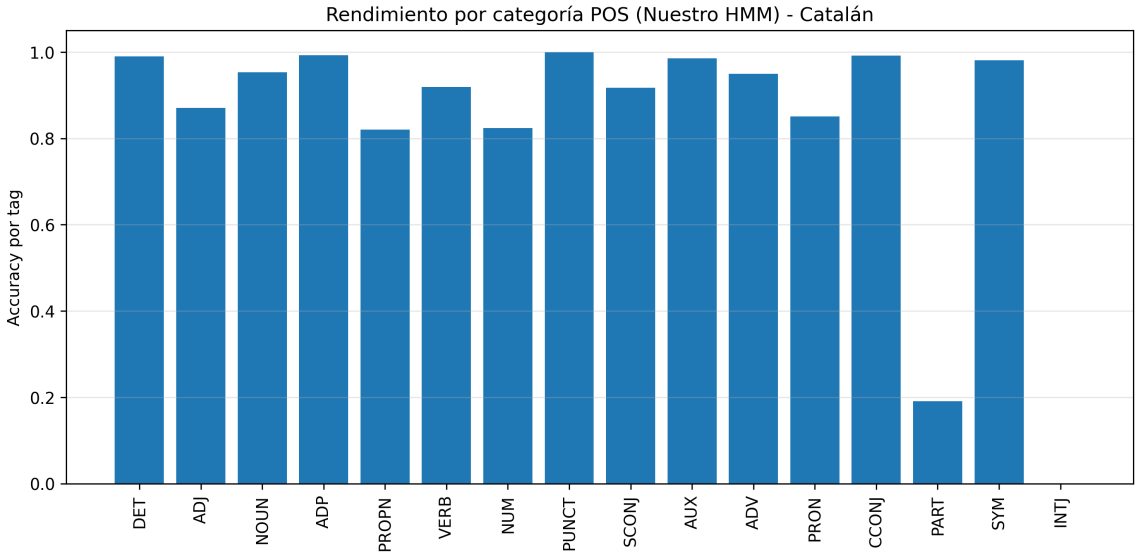

Figure 6: Per-tag accuracy of our HMM on the Catalan test set.

3.4 Qualitative checks

The notebook includes Viterbi decoding examples and joint-probability computations. Well-formed sequences (e.g., ADJ→NOUN→VERB) obtain higher joint probability than anomalous orders, echoing the "Colorless green ideas" contrast from class. Random samples drawn from the HMM also show plausible tag alternations, illustrating that learned transitions encode morpho-syntactic structure rather than semantics.

4 Conclusions