# Implementation of custom HMM for Basque and Catalan

Computational Syntax

Josu Bayer, Ane Paniagua, Ander Peña, Beñat Alkain

November 29, 2025

## Contents

**1 Introduction** **1**

**2 Methodology** **2**

**3 Results** **3**

3.1 Overall accuracy by split and language . . . . . 3

3.2 Comparison with n-gram baselines . . . . . 4

3.3 Per-tag analysis . . . . . 5

3.4 Qualitative checks . . . . . 6

3.5 Basque test metrics . . . . . 6

3.6 Catalan test metrics . . . . . 7

**4 Conclusions** **8**

## 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.

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`.

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.

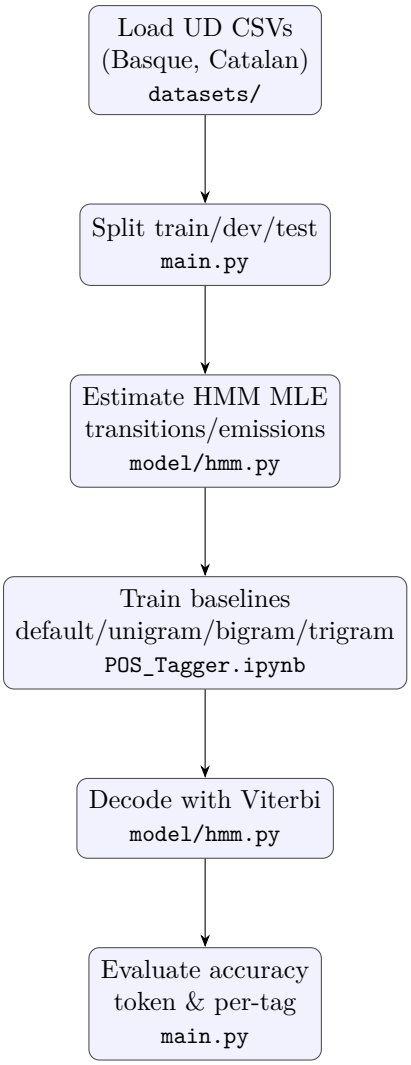


Figure 1: Methodological flow: data ingestion to evaluation with code pointers.

# 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. Numerically, for Basque we obtain 0.9569/0.8258/0.8189 (train/dev/test) with NLTK and 0.9693/0.8622/0.8560 with our HMM; for Catalan 0.9703/0.9453/0.9430 versus 0.9761/0.9477/0.9445. The dev–test gap is larger in Basque (e.g., 0.8622→0.8560) than in Catalan (0.9477→0.9445), reflecting higher sparsity from agglutinative morphology.

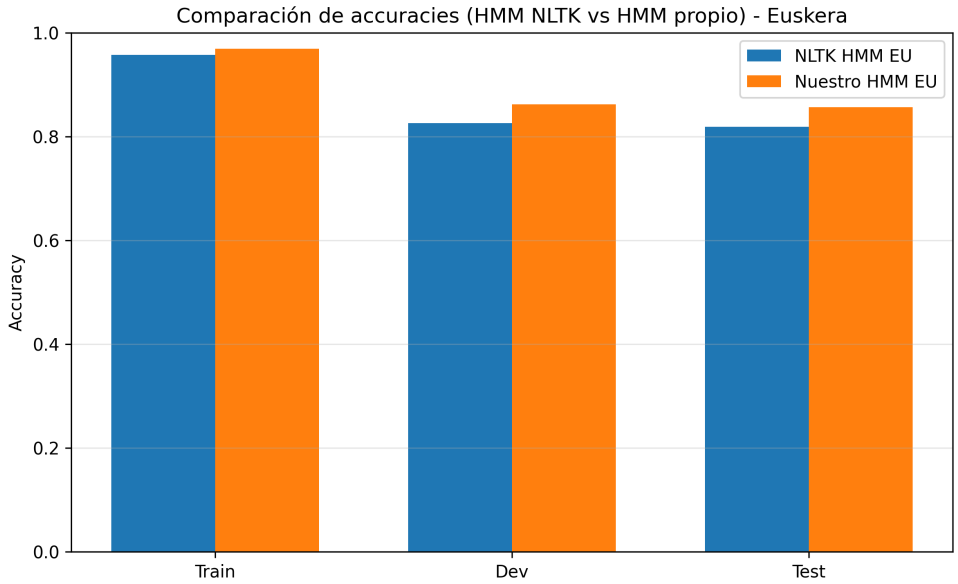


Figure 2: Basque: accuracy of NLTK HMM vs. custom HMM on train/dev/test.

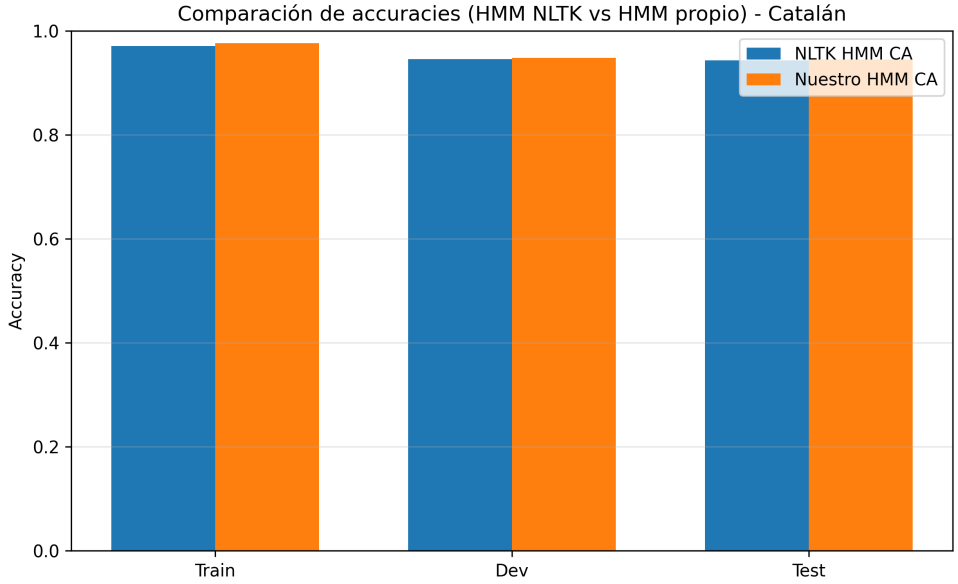


Figure 3: Catalan: accuracy of NLTK HMM vs. custom HMM on train/dev/test.

### 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. On the Basque test set the n-gram taggers reach 0.859 (unigram), 0.866 (bigram) and 0.866 (trigram), slightly above our HMM (0.8560) but ahead of the NLTK HMM (0.8189). For Catalan, n-grams reach 0.921/0.935/0.935, clearly below the NLTK HMM (0.9430) and ours (0.9445). Figure 4 visualizes these differences when combining plausible transitions (e.g., ADJ→NOUN) with emissions.

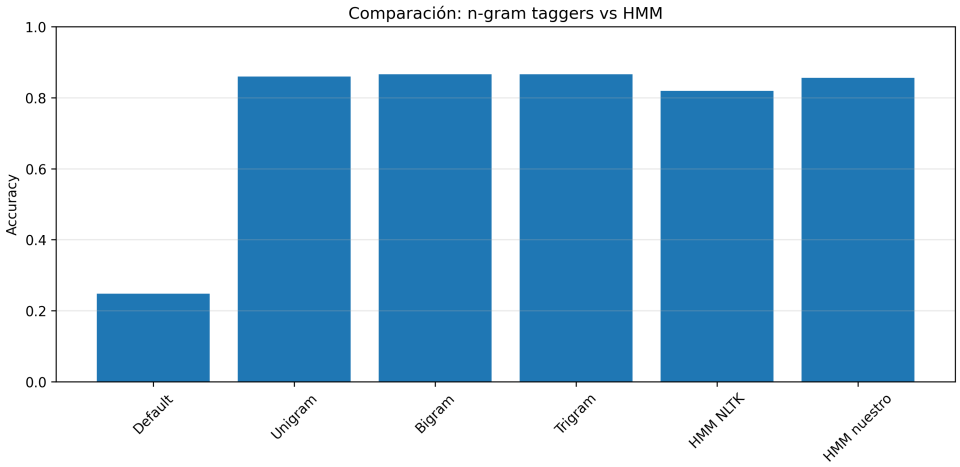


Figure 4: Token accuracy: default/unigram/bigram/trigram vs. HMMs (Basque split).

### 3.3 Per-tag analysis

Figures 5 and 6 detail per-tag accuracies for our HMM. Although the bars vary by tag, the quantitative pattern is consistent: closed classes (DET, ADP, AUX, PUNCT) cluster at the top, while open classes (NOUN, VERB, PROPN, ADV) drop. In Basque the drop is sharper due to larger vocabulary and sparser emissions; in Catalan the spread is milder.

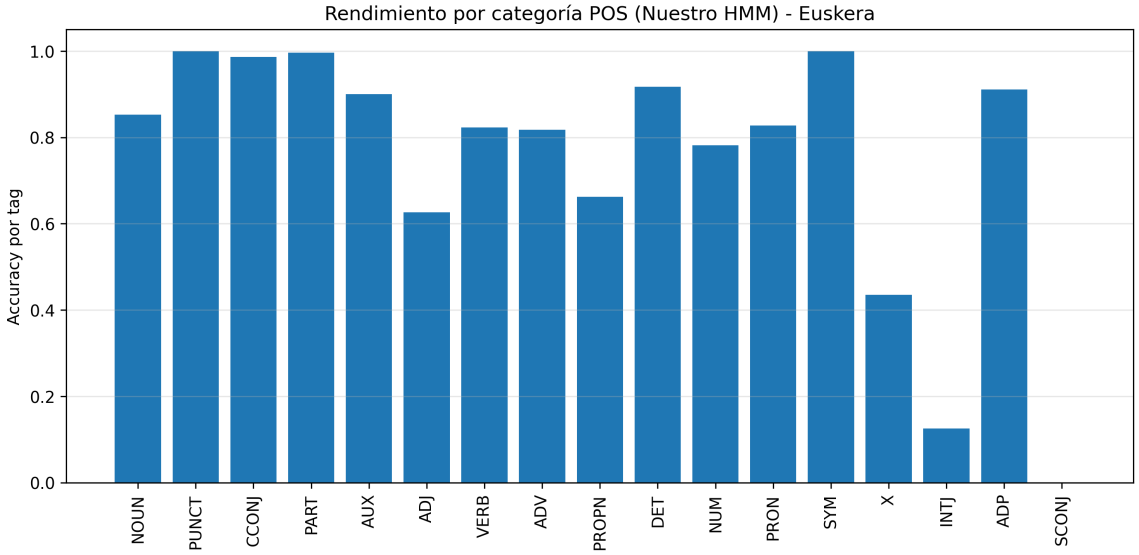


Figure 5: Per-tag accuracy of our HMM on the Basque test set.

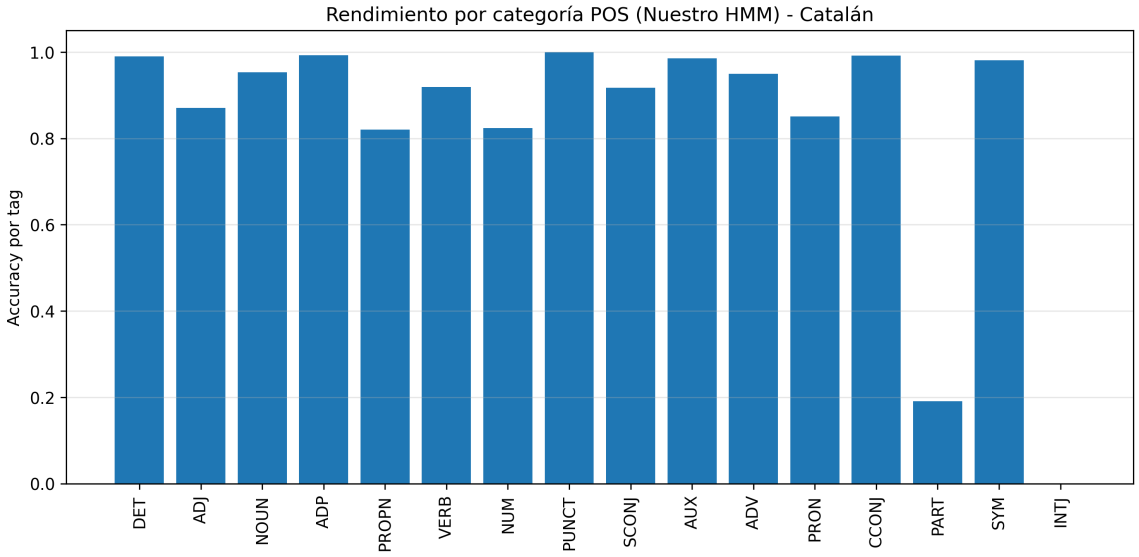


Figure 6: Per-tag accuracy of our HMM on the Catalan test set.

### 3.4 Qualitative checks

The joint probability for a Basque test sentence was $4.29 \times 10^{-13}$, consistent with multiplying probabilities over long sequences. Random samples from the HMM show plausible tag alternations (ADJ→NOUN→VERB), reinforcing the generative interpretation.

### 3.5 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 | | |

### 3.6 Catalan test metrics

Running `main.py` on the Catalan test set yields 0.94479 accuracy. Table 3 summarizes precision/recall/F1 by tag; macro-F1 is 0.836 (macro recall 0.827) and weighted-F1 0.944, showing higher overall performance and fewer drops on rare classes compared to Basque.

Table 3: Catalan test set: precision/recall/F1 by tag.

| Tag | Precision | Recall | F1 |
|---|---|---|---|
| ADJ | 0.906 | 0.870 | 0.888 |
| ADP | 0.980 | 0.993 | 0.986 |
| ADV | 0.927 | 0.949 | 0.938 |
| AUX | 0.949 | 0.985 | 0.967 |
| CCONJ | 0.988 | 0.992 | 0.990 |
| DET | 0.951 | 0.990 | 0.970 |
| INTJ | 0.000 | 0.000 | 0.000 |
| NOUN | 0.928 | 0.953 | 0.940 |
| NUM | 0.922 | 0.824 | 0.870 |
| PART | 1.000 | 0.190 | 0.320 |
| PRON | 0.907 | 0.851 | 0.878 |
| PROPN | 0.924 | 0.832 | 0.876 |
| PUNCT | 0.984 | 0.992 | 0.988 |
| SCONJ | 0.776 | 0.916 | 0.840 |
| SYM | 0.990 | 0.981 | 0.986 |
| VERB | 0.954 | 0.919 | 0.936 |
| **Macro-F1** | 0.836 (macro recall 0.827) | | |
| **Weighted-F1** | 0.944 | | |
| **Accuracy** | 0.94479 | | |

# 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: 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 different 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). In Catalan, overall scores rise (accuracy 0.94479, macro-F1 0.836): closed classes remain high (ADP 0.993 recall, PUNCT 0.992 recall), and even open classes improve (NOUN 0.953 recall, VERB 0.919 recall), with only INTJ and PART showing low recall (0.000 and 0.190).
- 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.