Lab 2 : Task 1 - Lemmatization

Idrissa Dicko Tyler Marino Simon Khan

January 22, 2026

Results and Comparison with spaCy

We evaluate three lemmatization approaches on the testing set:

- **MFL token → lemma**: a most-frequent-lemma baseline learned from the training data, without PoS information.
- **MFL (token, PoS) → lemma**: the same approach, but incorporating the part-of-speech tag to reduce lexical ambiguity.
- **spaCy baseline**: lemmatization produced by the `fr_core_news_sm` model, used as a general-purpose reference.

Overall Accuracy

Model	**Accuracy**
MFL token → lemma	0.9531
MFL (token, PoS) → lemma	**0.9653**
spaCy baseline	0.8150

Table 1: Comparison of overall accuracy between the proposed baselines and spaCy.

Both proposed approaches significantly outperform the spaCy lemmatizer, with an improvement of more than **15 accuracy points**. This result is expected, as the proposed models are directly trained on the same annotation scheme and domain as the evaluation data, while spaCy is a generic lemmatizer.

Known vs Unknown Words

To better understand the behavior of the models, we distinguish between words seen during training (*known words*) and unseen words (*unknown words*).

Model	**Known accuracy**	**Unknown accuracy**
MFL token → lemma	0.9694	0.6988
MFL (token, PoS) → lemma	**0.9855**	0.6828

Table 2: Accuracy breakdown for known and unknown tokens.

For known words, the accuracy is extremely high, reaching **98.6%** when PoS information is used. This shows that lemmatization is almost deterministic for observed lexical forms and that PoS tags effectively reduce ambiguity for homographic tokens (e.g., noun vs verb forms).

For unknown words, performance drops to approximately **69%**, highlighting the intrinsic difficulty of generalizing to unseen forms. Interestingly, incorporating PoS information does not improve performance on unknown words and even slightly degrades it, since PoS-specific token–lemma pairs remain unseen in these cases and the prediction relies mainly on the fallback strategy.

Discussion

Despite its simplicity, the most-frequent-lemma approach proves to be a very strong baseline for lemmatization. The inclusion of PoS tags provides a clear benefit for known words, while the main limitation of the approach lies in its handling of out-of-vocabulary tokens. Overall, the proposed methods demonstrate that simple supervised lexical strategies can outperform more complex general-purpose lemmatizers when the domain and annotation scheme are well matched.