Q1:

The majority issue of tokenization is OOV, it recognizes the new word as PROPN in this system. *Left* is usually an adjective in the training corpus. The training corpus defines tokens LAS and LDWS as PROPN, it should be included. The system does not include a language detector because there are Spanish words in the sentences and it is recognized as PROPN, “comprar” should be a verb in Spanish. All the unseen and new words that are not in the training corpus are proper nouns. I do not think the letter case affects the system. This system cannot analyze other words but English.

Q2:

To handle a better on unseen words, we can replace every stop-word in the training corpus.

Extra Credit:

The accuracy will be a little bit higher for seen words than unseen, the best way I predict tags for unseen words is to define them as PROPN.

The NOUN will be the easiest to predict and ADJECTIVE or VERB will be harder to predict because NOUN is always for name, place, or someone. Prepositions are easy to tag.

Prepositions are only capitalized if they are used adverbially, we should capitalize when the tokens with more than 5 characters.