It would make sense to create a language detector on the character level instead of at the word level because there will be higher grouping of letters together in one language than another. Especially if you are using 4-grams or bigger ones there should be patterns of how the letters are grouped. For instance, in Spanish, because of the way words are conjugated, there will be more words ending in a specific pattern as many different words will be conjugated with the same ending. The same goes for English, there are many words that are conjugated to end with something like '-ing' or '-ly' more so than Spanish.

For the Spanish documents, the beginning of them are in English because they are from Project Guttenbrug, so it says all of the stuff they say before every book. Besides that, it seems they remain in the one language; the English ones seem to be entirely in English.

For English it was able to predict with as few as around 150 tokens, but for Spanish at that amount it was predicting English. The tipping point for predicting Spanish was a little over 300 tokens.

The first thing that came to mind is comparing accuracies, as in see which one gets more correct, but with the second part of the question, where they both have 100% accuracy, that changes my answer. I'm thinking you could try to start seeing which one gives the right answer with the fewest amount of data input. You could also go off of time, if one takes an hour, but the other takes 10 minutes, then the faster one would be better. So faster as well as being able to take a smaller amount of data and still get the right answer would help in deciding the winner.

When training with very few tokens, such as 100-300, the accuracy is very low, probably actually around 0 with the correct predictions being lucky. As it increases to using one of the pages that is in the training/en or training/es folders then that is enough for it to get 100% accuracy. Once it got to 1000 tokens, then it was back at 100% accuracy again. So a graph would look something like accuracy around

0 for 100 tokens and slowly approach 100% accuracy as it got to 1000 tokens then begin to taper off and not increase much past that.

It predicted that the language is in Spanish which is what I was expecting. I believe the reason it predicts Spanish for French is because both are a Romance language and so have very similar roots and thus likely have a lot in common. I don't know much about French, but I imagine the conjugations might be somewhat similar as well as smaller words 'es'.