COMP 472 – Mini project 3

Language models - Analytical report

December 4th, 2018

Victor Debray (ID: 40102554)
Theo Penavaire (ID: 40102474)

This report provides a description of the research made for the COMP 472 course's third project. For this project we had to build three language models and experiment with language identification.

We chose to use C++ to develop our program. Being a low-level programming language, we felt it is a good choice for the high computation involved in the language model training.

In this report, we will first present our experimental setup. We will then study the resulted output and draw final conclusions.

The subject asked us to work with english, french and one our choosing. Each corpus that will be used to train the language model has diacritics removed. We picked Euskara (language spoken in Victor's home region of France, the Basque Country). One convenient aspect of this language is that it does not use any accents, besides the \`ñ\` that can be easily replaced by the \`n\`.

I. Architecture

At the program's input the user provides the three corpus and the file containing the sentences seperated by a return character. The goal of the program is to identify the language of each sentence. With the corpuses, are trained the corresponding language's language models.

A. Uni-gram

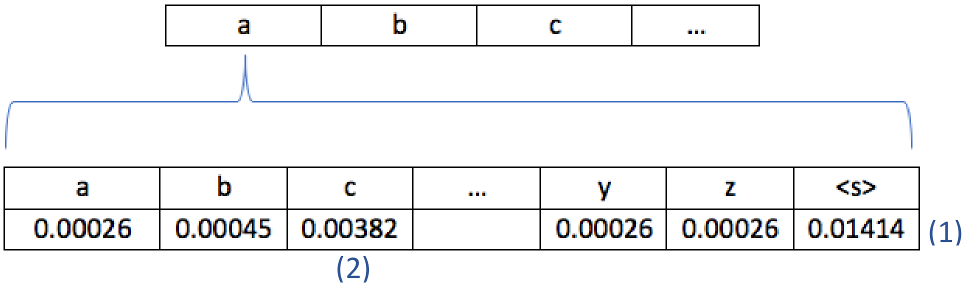
A uni-gram is a character. We compute with the uni-gram the weight of a character in a language. The uni-gram is represented as follows.

a	b	c	…	x	y	z
0.0826548	0.0101307	0.0284984		0.0053967	0.0017989	0.001230

French smoothed uni-gram frequencies

B. Bi-gram

The bi-gram, a sequence of two characters. We compute the weight of a bi-gram by taking the reference character and the one before it (the history). We also implemented the case when a character is the first of a word. So it is represented as a bi-dimension array of 26 by 26 + 1 (for the probability of being the first letter of a word).



a	b	c	…

a	b	c	…	y	z	<s>
0.00026	0.00045	0.00382		0.00026	0.00026	0.01414

Euskara smoothed bi-gram frequencies. P(a|<s>) would correspond to the probability in cell annotated (1). P(a|c) the cell (2).

Since the lookup is done by a one size key (a character) the model is structured in a array. We bring the value of the character to a 0 to 26 index

C. Result

The sentences, are then compared to each model. For each gram (uni-gram and bi-gram) of the sentence we compute the additioned log probability taking the value from each language's language model. The language model that has the closest result identifies the language of the sentence.

II. Classification of 30 sentences

We were asked to classify 30 sentences with our language models. Most of the time (70% for unigrams, 73% for bigrams), the sentences were correctly classified. It is interesting to know why the other weren't.

- *Welcome Jean Baptiste !*
- *Bienvenue William !*

We think that for the two first sentences, the language models were tricked by the names. "Jean-Baptiste" is a french name in an english sentence and "William" is an english one in a french sentence. They both contain unigrams and bigrams that are very common in their language. As the sentences are very short, it is enough to trick the Language Model and make it classify the sentence in the wrong category. This is clearly visible when comparing the final sum of the log probabilities for the bigrams (for "Welcome Jean Baptiste !"):

FRENCH: P(e|t) = 0.00845346 ==> log prob of sentence so far: **-173.027**

ENGLISH: P(e|t) = 0.00578176 ==> log prob of sentence so far: **-173.513**

EUSKARA: P(e|t) = 0.0145563 ==> log prob of sentence so far: -198.762

We can see that they are very close between french and english, and that changing a single letter could change the whole classification.

- *Go !*
- *Non ?*

These are very short sentences, and most important, ambiguous ones. Indeed, "Non ?" is the euskaran question for "Where ?", but could also be interpreted correctly in french and english (even if this wouldn't really be a sentence). The interesting fact is that our unigram Language Model classified it as french, and our bigram as english. None of them classified it as euskara. "Go !" was classified as euskaran, mostly because of the more important frequency of the 'G' letter than in other languages (see the Occurrence frequencies graph next chapter).

III. Going further

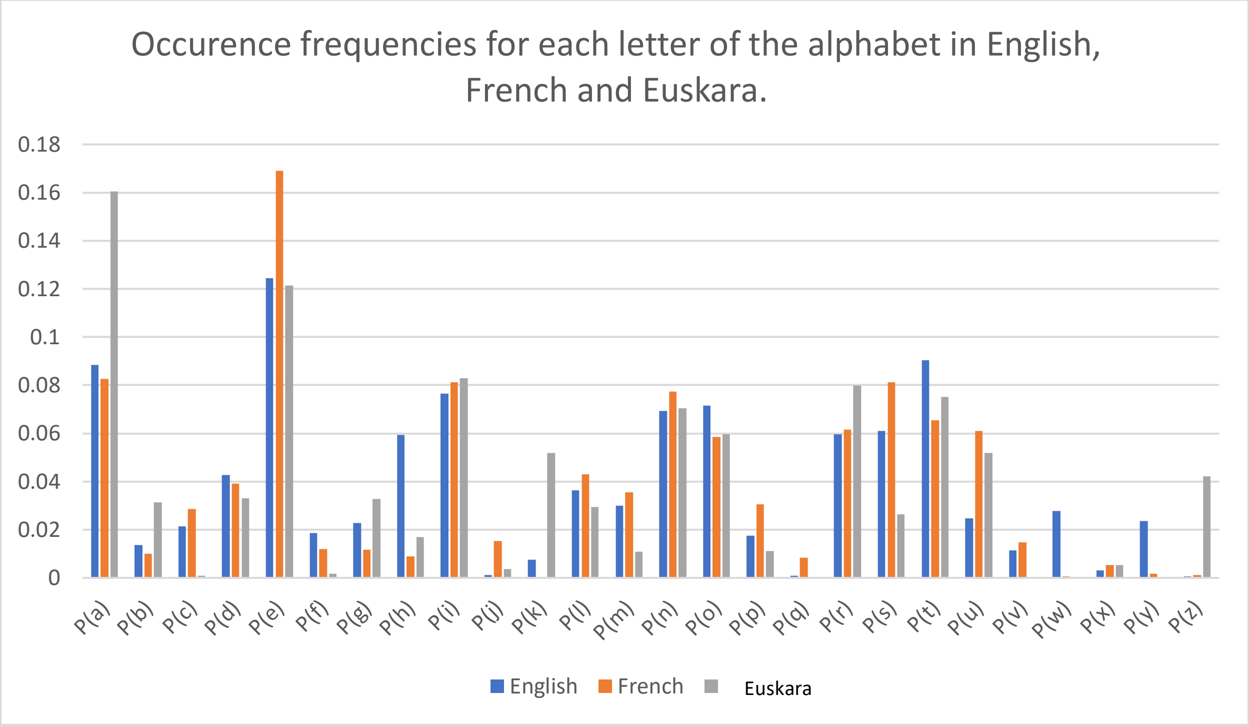
Language identification finds most of its applications in language translation.

But studying language models and language identification is also a great way to understand what makes a language unique among others. Indeed, by classifying languages, we gather data that can be organized into features (n-grams for instance). Analyzing these features can lead to observing similarities and differences between languages. This can be of a great help for linguists and historians !

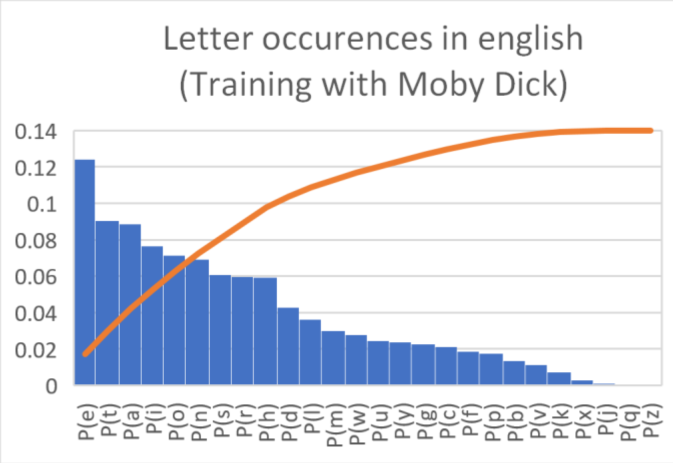
This is what motivated us to try to differenciate the three languages, according to some features. In this section we will therefore try to establish a list of key similarities and differences between each language.

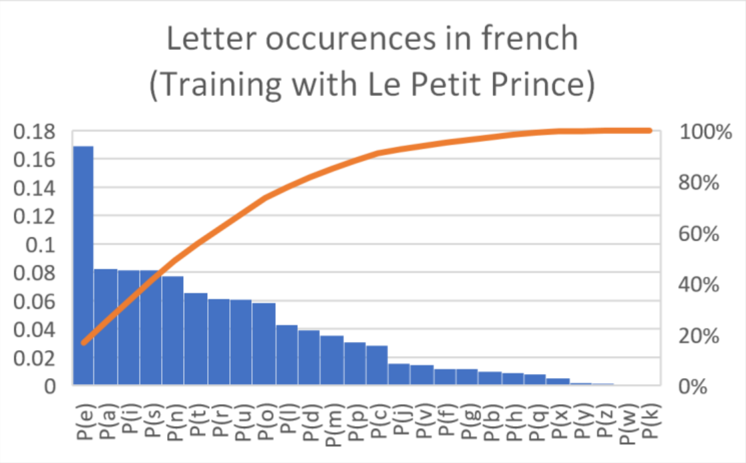
A. Unigrams

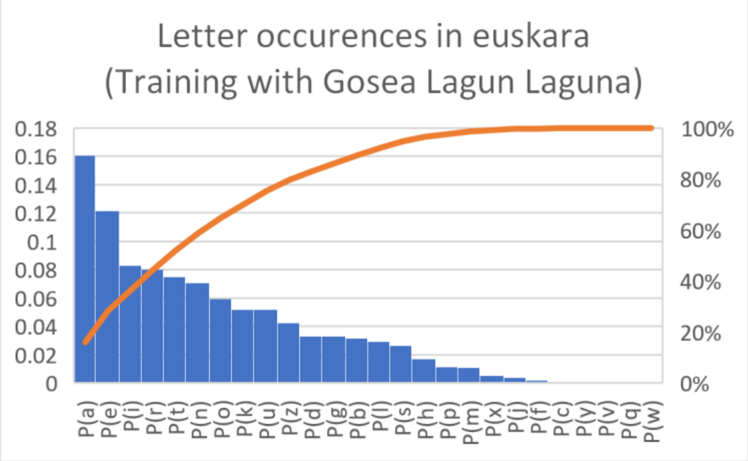
We first summarized the occurrence frequencies for each letter for english, french and bask.



Overall, for a given letter, the occurrence frequencies are similar for the three languages. English and French have common roots however bask is a language of itself, having some few words influenced by arabic and latin for the new technical words. French and english were influenced by the latin and the indo-european languages. Also we can see that vovels are more commonly used in all three languages than consonants. At least this is true for ‘a’, ‘e’ and ‘i’.







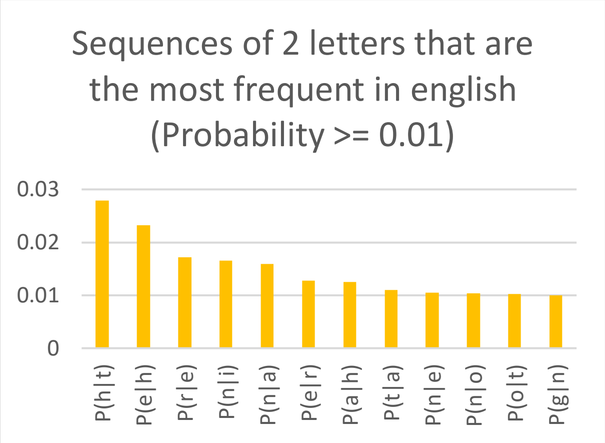
There are some subtle differences between these languages though:

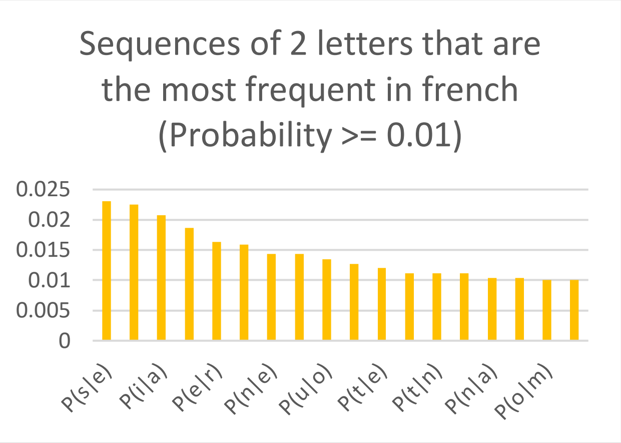
- ‘a’, ‘k’, ‘z’ are more common letters in bask than in other languages.
- ‘e’, ‘p’, ‘j’, ‘q’ are more common letters in french than in other languages. The letter'e' is even twice as used as the second most used letter (‘a’) ! This could be why the french language uses so much accents (each variation of the ‘e’ letter with an accent could be considered as a different letter after all: pronunciation and meaning are different).
- ‘h’, ‘w’, ‘y’ are more common letters in english than in other languages.

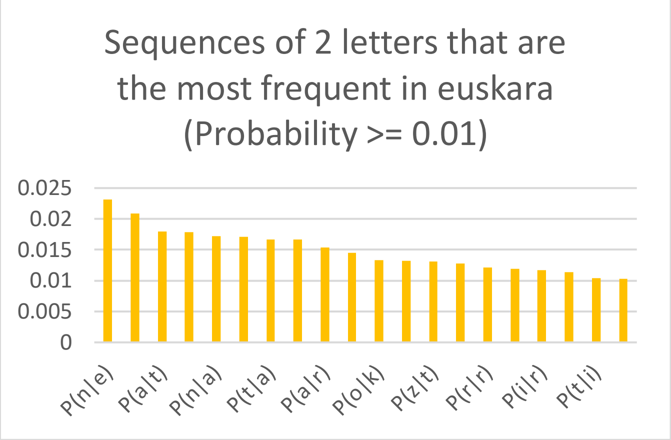
We could say that these letters form the “signature” of each language.

B. Bigrams

Following this idea of “signatures” for each language, we considered bigrams.







Most used sequences of two letters:

- “th”, “he”, “er”, “in” in english.
- “es”, “ai”, “re”, “en” in french.
- ‘’en’’, ‘’ta’’ , ‘’an’’, ‘’at’’ in euskara.

As a conclusion we can say that it is very easy to find key elements that define a language, and to compare them with other languages.

With more time and resources it would have been very interesting to pursue these experimentations, and maybe draw a bigger picture (common roots, history of languages…). This however begins to be out of the scope of this project.