# **Accepted Sentences**

* Le cheval dort.
* Napoléon représente les vieux à la chaise.
* Les poules chantant un chanson.

# **Rejected Sentences**

* C’est la vie.
* Je pense donc je suis.
* Je ne m'enfuis pas je vole.

# **Advantages**

The advantages of CFG French modelling are:

1. It can handle difficult lexical category system, distinguish masculine from feminine. It can help with generating sentences with proper grammar and find the right tagging.
2. When applied in translation, it can help translate certain words from languages which doesn’t feature masculine and feminine, and from those words supplement the whole sentence.

# **Disadvantages**

The disadvantages of CFG French modelling are:

1. It highly requires the standardization of the sentences. If wrong grammar or spelling appears in input data, it can make mistakes.
2. There are so many categories of non-terminals and when the CFG model is not well designed, there can be problems in recursive process. For example, noun phrases can include prepositional phrases and vice versa. The time complexity can also be huge.
3. In some conditions such as POS tagging, the context can have huge influence on the languages and tags, or two languages are mixed used. CFG model can fail on these conditions where context has huge influence.

# **Aspects which my CFG cannot handle**

This CFG model definitely can’t handle tense which is also complex in French. In addition, the phrases of proper nouns, passive voice, possessive pronouns and possessive adjectives are not considered in this model. It can also fail in question sentences， imperative sentences and inverted sentences.

|  |  |  |  |
| --- | --- | --- | --- |
| Improve/Data Set | Cipher\_1 | Cipher\_2 | Cipher\_3 |
| None | 1.12% | 0.0% | 16.17% |
| Laplace | 97.68% | 83.83% | 22.18% |
| Improved Text | 100% | 74.38% | 16.60% |
| Both | 100% | 74.38% | 16.60% |

# **Conclusion**

The accuracy values of different conditions on different datasets are shown above. Without using Laplace Smoothing the accuracy values on all three cipher sets are extremely low, closing to 0 especially in cipher\_2 and cipher\_3. When Laplace Smoothing is applied, the accuracy of these two data sets dramatically increase, reaching 97.68% in cipher\_1. However, the accuracy only increases around 6% in cipher\_3. I have experimented with another estimator which is LidstoneProbDist and the accuracy reach 98.02% and 84.78% respectively, but the performance on cipher\_3 is similar to Laplace.

Then I import a dataset of 20k sentences, transform them to lowercase, segment them into sentences, remove the characters not in 29 characters and extra spaces. When applying only those improved plain texts modelling, the accuracy values also dramatically increase in cipher\_1 and cipher\_2, while cipher\_3 remains similar. By applying both Laplace smoothing and Improved plain text modelling, the performance stays the same on all three datasets.

For cipher\_1, the cipher method is simple and simplex. So, within the small training datasets, the HMM algorithm can perform well and improved by additional plain texts. However, cipher\_2 and cipher\_3 apply random cipher method and the datasets are still small, which lead to poor performance of the HMM model because of small sample size. Compared to improved plain texts modelling which include 20 thousand additional sentences, the original datasets only consist of 13 sentences. As a result, improved plain texts modelling are more ‘powerful’ on the performance of models than Laplace smoothing and when applying both, improved plain texts modelling has the absolute influence because the 13 sentences of training datasets can be ignored.