# Data Sets

## Nature of the Data Sets

The training data as well as the test data sets have been compiled and anonymized by reputable institutions. It can therefore be assumed that the data offered by them is reliable and accurately reflects the French language.

As to the nature of the data sets, they depict a non-standard form of the language that is commonly found in chats and text-based communication such as SMS, Whatsapp, etc. This carries with it the complications posed by incorrect spelling, abbreviations, emoticons, etc.

## Training Data

As is per the norm in computational linguistics, the data of the program will be split into two main categories: test data and training data. The corpora are sourced from two main sources *A collection of online auction listings from 2005 to 2018 (anonymised)*(Gerstenberg & Hewett, 2019)and *CoMeRe Repository: Corpora of Computer-Mediated Communication in French* (Chanier,T. et al. 2014)

Seeing as how the data set from the auction listing is relatively small and already partially tagged, it is suitable to be the first half of the training data set. The other half will come *Etiquetage morpho-syntaxique du corpus FAVI [corpus]* Riou, S. & Sagot, B. (2016), which, while somewhat larger, has also already been partially tagged.

A training corpus can also be automatically generated using regular expressions

## Test Data

The test data will be sourced from *Corpora of Computer-Mediated Communication in French* (Chanier,T. et al. 2014) as they have other data sets that are untagged and also depict chat-based communication.

# Tag Sets

There are multiple tag sets that are universal, in the sense that they like those proposed by Penn State or a *universal part-of-speech tag set* (Petrov et al., 2012) do their best to be representative of the features that are present in all languages.

Then there are French-specific models such as the French TreeBank (FTB, Abeillé et al; 2003), *28-tag tagset* (CC tagset, Crabbé and Candito, 2008) and Treetagger (Schmid, 1994).

## Universal

I will be using two main tag sets, one universal (Petrov et al., 2012) .

|  |  |
| --- | --- |
| VERB | verbs (all tenses and modes) |
| NOUN | nouns (common and proper) |
| PRON | pronouns |
| ADJ | adjectives |
| ADV | adverbs |
| ADP | adpositions (prepositions and postpositions) |
| CONJ | conjunctions |
| DET | determiners |
| NUM | cardinal numbers |
| PRT | PRT |
| X | other: foreign words, typos, abbreviations |
| . | punctuation |

## French

The second, a French-specific tagger that will be an minor extension of Treetagger (Schmid, 1994) to better accommodate and capture the linguistic features of chat data.

|  |  |
| --- | --- |
| ABR | abreviation |
| ADJ | adjective |
| ADV | adverb |
| DET:ART | article |
| DET:POS | possessive pronoun (ma, ta, ...) |
| INT | interjection |
| KON | conjunction |
| NAM | proper name |
| NOM | noun |
| NUM | numeral |
| PRO | pronoun |
| PRO:DEM | demonstrative pronoun |
| PRO:IND | indefinite pronoun |
| PRO:PER | personal pronoun |
| PRO:POS | possessive pronoun (mien, tien, ...) |
| PRO:REL | relative pronoun |
| PRP | preposition |
| PRP:det | preposition plus article (au,du,aux,des) |
| PUN | punctuation |
| PUN:cit | punctuation citation |
| SENT | sentence tag |
| SYM | symbol |
| VER:cond | verb conditional |
| VER:futu | verb futur |
| VER:impe | verb imperative |
| VER:impf | verb imperfect |
| VER:infi | verb infinitive |
| VER:pper | verb past participle |
| VER:ppre | verb present participle |
| VER:pres | verb present |
| VER:simp | verb simple past |
| VER:subi | verb subjunctive imperfect |
| VER:subp | verb subjunctive present |

The reason for using two data sets is for the sake of comparability. By having one tag set that is universal and the other that is French-Specific, the reliability of the tagger with respect to universal and specific taggers can be assessed and determined.

# Pre-Processing

## XML

As is typical of linguistic data, this data is saved in an XML format that must be properly passed so that the sentences and tokens can be obtained.

## Tokenization

Before the data can be properly assessed and tagged, it must be tokenized into the smallest possible linguistic units. This will be done using naïve and advanced heuristics.

# Python Libraries and Modules

## Standard

## Pip

## Custom

# Algorithms

# References