Syllabification

Phonetics Project

by

Eduardo Calò, Thibo Rosemplatt

under the guidance of

Katarina Bartkova Jouvet

December, 2019

Contents

Chapter 1

Linguistic approach to syllable

1.1 General

Among many different definitions of syllable, on the basis of which linguistic theory is taken into account, the most widely accepted is that a syllable is a linguistic unit of organization for phonemes built around a peak of sonority. A typical syllable can be structured into 3 main components: *onset*, *nucleus*, *coda* (last two components can be grouped into *rhyme*). Among these 3 parts, the nucleus is the only one which is mandatory, with the onset and coda possibly being empty (see Figure 1.1).

The syllable is regarded as a fundamental linguistic unit, and this is highlighted by the fact that speakers seem to be able to syllabify words without effort. However, one of the main problems in phonology is to theoretically formulate rules for finding boundaries between adjacent syllables. In order to solve this problem, one of the

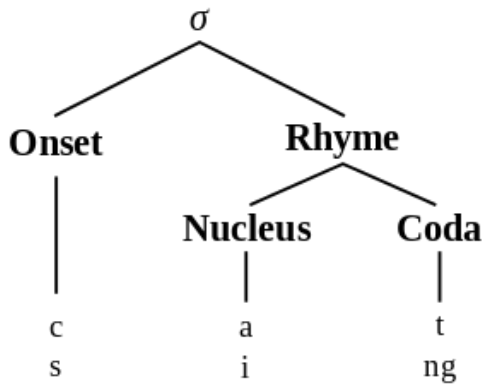

Figure 1.1: Typical syllable structure.

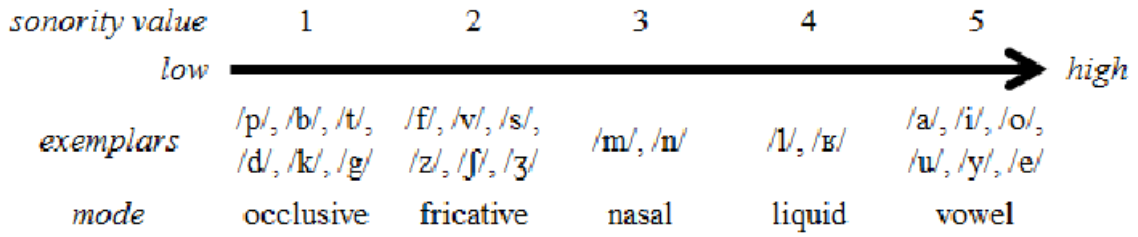

Figure 1.2: Sonority hierarchy.

most influent theories proposed by scholars is the sonority hierarchy. It consist of a scale in which phonemes are divided into different classes on the basis of loudness. Usually the manner of articulation of a sound is used as a baseline for creating this scale (see Figure 1.2). The louder the sound, the more likely it is that that sound can appear as syllable nucleus. Usually the position of phonemes composing a syllable follow this scale, with the consonants (least intense) taking onset or coda positions, and vowels (most intense) as nucleus.

1.2 French syllable

French syllable is no exception. Each syllable contains a nucleus which is composed obligatorily by a vowel, followed by one or several optional consonants in coda position. The nucleus can be preceded by one or several consonants in onset position. That being said, a vowel is the clear indicator of the presence of a syllable, so the starting point for counting syllables in French are the vowels. For example, if 3 vowels are present in a word, then that word is trisyllabic. To place syllabic boundaries, phonotactic rules, limiting how phonemes can be combined together, come now into the equation. These rules depend on the number of consonants between 2 vowels, and on the class of these consonants.

In addition to the already given rules, we added another rule to handle the "floating s" between two syllabic structures, in cases like /apstRakt/. Here the syllable boundary must be between the consonants other than liquids, and we have decided to put the boundary after the fricative: /aps-tRakt/.

Chapter 2

Automatic syllabification

This chapter aims at giving an overview on how the program works to handle the syllabification process, and provide the output file.

2.1 Code pipeline

1. The **process_file** function is launched. This function will read the *Input_file.txt* file, and clean it to prepare a good processing. Then, each line of the file will be formatted via the **format_output** function. Each line passed as input in this function is processed as follows.

2. The line processed contains the orthographic and phonetic forms of a French word. The **format_output** function takes them as input and returns:

   - The orthographic form *(Already known)*
   - The orthographic consonant/vowel (CV) form
   - The phonetic form *(Already known)*
   - The phonetic consonant/vowel (CV) form
   - The phonetic form, split by syllables with hyphens
   - The CV phonetic form, split by syllables with hyphens

   The most interesting items regarding this project are the last two, in particular, the penultimate. The phonetic splitting by syllable is the core of the project. This is done via the function **syllabic_phonetics**.

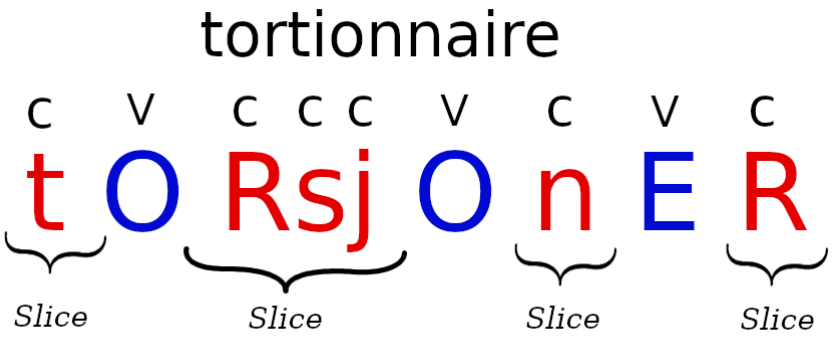

Figure 2.1: Visualization of the *slices* within a phonetic form.

3. The **syllabic_phonetics** function takes a phonetic form called *sampa* as input. As we noticed that the useful information for the syllabification process is determined by the characteristics of the consonants in between two vowels (see Section 1.1), the first thing we did is to retrieve the indexes of the vowels of *sampa* via the function **spot_vowels**. *Sampa* will then be split into slices of consonants surrounded by vowels and/or by the *sampa* boundaries (see Figure 2.1). Then, the position of the hyphen in each slice will be computed following the rules mentioned in Section 1.2. At this point, the position of the slices within *sampa*, and the position of the hyphens within the slices are known. The final step of the syllabification process is to deduce the position of the hyphens inside the whole *sampa* itself. This is computed simply by adding the position of a slice within *sampa* with the position of the hyphen within this slice. Doing it for each slice, we have the position of the hyphens inside sampa phonetic transcription.

4. The last step is to write the result inside *output_file.txt*. The preceding process is repeated for each line of the cleaned *Input_file.txt*, and the whole *output_file.txt* is created.

2.2 Special cases

2.2.1 Multiple pronunciations

Some lines of the input file contain a semicolon inside the sampa transcription:

```
1397    concupiscent k1kypis@;k1kypis
```

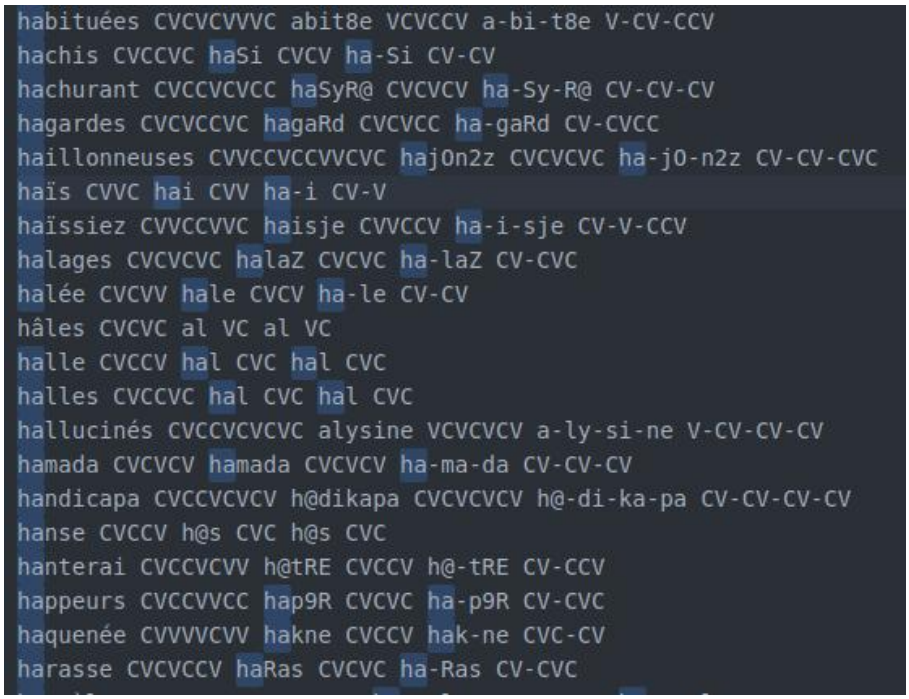

Figure 2.2: Examples of the h phoneme present in the input file.

This is because some orthographic words can have multiple pronunciations. Such cases are few, but they had to be handled for a good performance of the program. Thus, in the cleaning process mentioned in Section 2.1, such input lines were transformed into two lines, each containing one of the sampa forms:

```
1397    concupiscent k1kypis@
1398    concupiscent k1kypis
```

2.2.2 H

We noticed that some sampa transcriptions of the input file contain the *h* phoneme. This is surprising because this phoneme does not exist in the French language.
An explanation to this phenomenon is that the *h* here may be misused. It seems that its function here is to represent a break of the *"liaison"*, meaning that when this *h* is present, the last phoneme of the previous word that it usually pronounced will be silent. This is confirmed by the multiple examples we found (see Figure 2.2), which are known to break the French *"liaison"*. This is what we assumed, and we did not specifically handled this case.

Chapter 3

Most frequent syllables

3.1 Code overview

The goal here was to to retrieve the 15 most used phonetic CV-forms, macro-classes, and phonetic syllable from the output file. The first step has been to obtain 3 distinct flat lists, each containing all the different occurrences of each item aforementioned. This has been performed with the functions **get_all_cv_forms**, **get_all_macroclass_forms**, and **get_all_plain_syllables**.
Once the lists have been collected, the library *Collections* has been used to sort them by number of occurrence. Thus, the most used items and their number of occurrence have been retrieved. They are displayed, for each category, in Figures 3.1, 3.3, 3.5.

3.2 Comments on the results

The most striking phenomenon is that the distribution of CV patterns is very irregular. As shown in Figure 3.2, the first elements capitalize the vast majority of occurrences, and the following ones appear few times.
On the other hand, the macro-class forms and the phonetic syllables forms are distributed in a smoother way (see Figures 3.4, 3.6).

CV form	Number of occurence
CV	10968
CVC	3112
CCV	2418
V	1672
CCVC	587
CVCC	381
VC	306
CCCV	235
CCCVC	79
VCC	75
CCVCC	34
CVCCC	8
CCCVCC	8
CCCCVC	4
CCCCVC	3

Figure 3.1: Results for the most used phonetic CV forms.

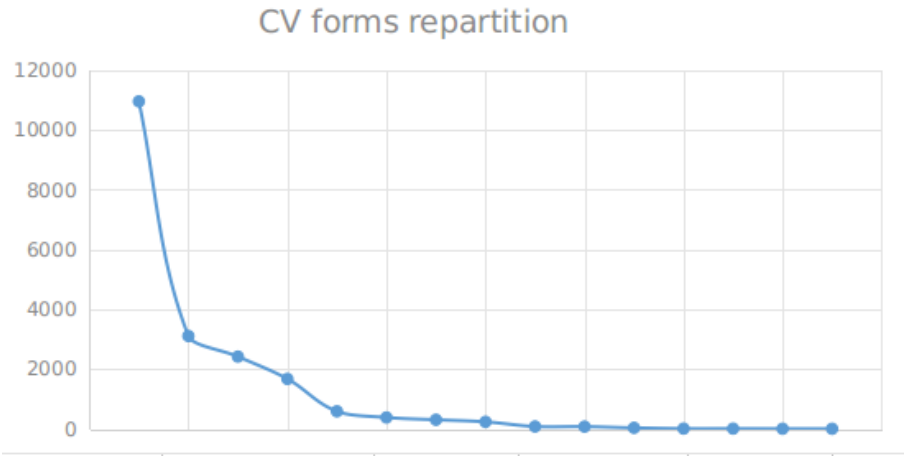

Figure 3.2: Distribution of the most used phonetic CV forms.

Macroclass forms	Number of occurence
Stop unvoiced, vowel	2752
liquid, vowel	2099
nasal, vowel	1768
vowel	1672
Fricative unvoiced, vowel	1565
Stop voiced, vowel	1524
Fricative voiced, vowel	1043
Stop unvoiced, liquid, vowel	827
Stop unvoiced, vowel, liquid	525
stop voiced, liquid, vowel	425
fricative unvoiced, semi-vowel, vowel	308
fricative unvoiced, vowel, liquid	226
semi-vowel, vowel	215
nasal, vowel, liquid	202
stop voiced, vowel, liquid	170

Figure 3.3: Results for the most used macro-class forms.

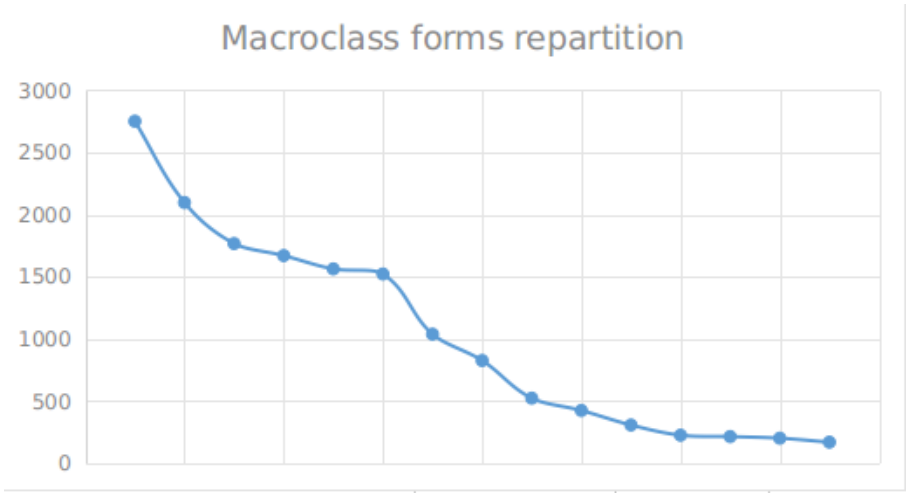

Figure 3.4: Distribution of the most used macro-class forms.

Phonetic syllable	Number of occurence
a	470
de	458
e	334
te	330
ti	282
@	272
Ra	267
Re	265
m@	254
li	235
k1	227
5	224
R*	223
se	193
si	192

Figure 3.5: Results for the most used phonetic syllables.

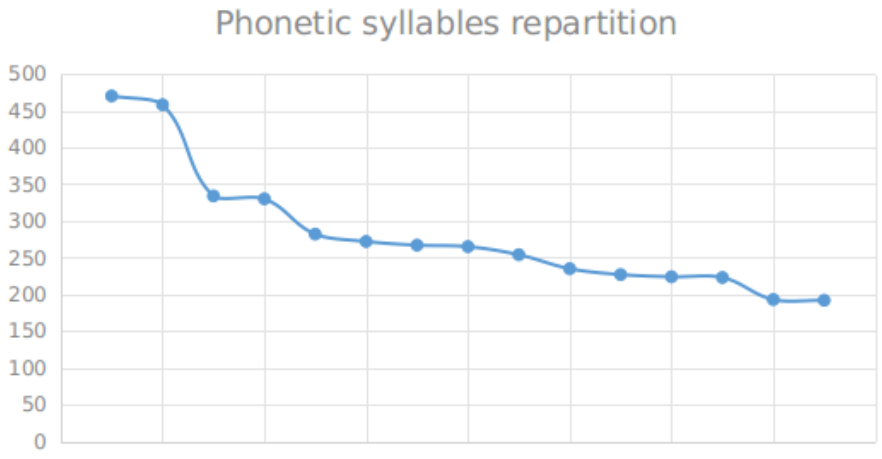

Figure 3.6: Distribution of the most used phonetic syllables.