The Metrometer: a Tool for Analysing French Verse

VALÉRIE BEAUDOUIN
CREDOC and EHESS, FR.

FRANÇOIS YVON
TELECOM PARIS, FR.

Abstract

In this article, we present the "metrometer", a computing tool, capable of identifying, in any kind of French text input, metrical components, that is syllables. The choice of the syllable is very natural: French poetry is known to be mainly syllabic. More specifically, our software performs a complete transcription of the input into the corresponding sequence of phonemes, using a transcribing module consistent with the specific phonology of classical poetry, which allows the marking and counting of metrical syllables. Our system outputs as well with a series of useful markers (syntactic tags, word boundaries, etc.). This tool has been developed and tested on a corpus containing all P. Corneille’s and J. Racine’s plays, totaling about 80,000 verses, and has proved to be almost error free on these data. The metrometer shall provide a very effective tool to tackle the study of rhythm in French poetry with quantitative methods.

1 Introduction

The computer tool, the metrometer, presented below has been developed with the specific purpose of exploring, on large verse corpora, the various surface aspects under which rhythmic phenomena appear. We have primarily limited our scope to a very restricted and rigidly codified form of verse, namely the classical (XVIIth century) alexandrine. More precisely, this tool has been developed and tested on a corpus containing all P. Corneille’s and J. Racine’s plays, totaling about 80,000 verses, and has proved to be almost error free on these data. The metrometer shall provide a very effective tool to tackle the study of rhythm in French poetry with quantitative methods.

2 System overview

While working on phonemic transcription problems, the need for a complete and accurate parsing of the input text progressively emerged. Such an analysis is especially useful to disambiguate heterophon homographs, and to address in detail the problem of liaison in the verse. At the current stage of our work, a simple part-of-speech tagger would have been enough to solve most of our difficulties; it is however obvious that a full syntactic analysis would eventually be a necessary step for any serious analysis of rhythm. This explains why our transcription module corresponds to an independent layer on top of the syntactic analysis system developed at TELECOM PARIS by P. Constant (1991).

We shall present the design of this module, before explaining how text strings are converted into phoneme sequences.

2.1 Syntactic analysis
P. Constant's syntactic analyser radically differs from most known parsing systems of French. It relies upon the strong assumption that syntactical ambiguity is a major property of natural languages, and that, consequently, ambiguities should not be regarded and treated as accidental events. His system therefore offers an effective and structural way to handle ambiguities.

The second unusual aspect of his model is that the concept of a "well-formed" sentence, a central element of the theory of formal grammars is not relevant. On the contrary, the system is able to produce a parse for any kind of sentence, even incorrect ones. In this respect, it is robust to local orthographic "catastrophes", such as in "Un petit chats" (a small cats). This proved to be for us a valuable feature, for one main reason: our maximal analysis unit being the verse, we needed a system capable of processing such units, which do not always correspond to complete sentences. There is of course another good point in using such a system: orthographic typos, also present in our corpus can safely be ignored.

The analysis proceeds by applying sequentially a series of heuristics, each heuristic or group of heuristics functioning as an independent layer. The first layer concerns text segmentation and lexical heuristics functioning as an independent layer. The series of heuristics, each heuristic or group of such heuristics (a layer) is dedicated to the structuration of verb groups, etc. The reduction of the number of possible parses goes together with the refinement of the syntactic structuration of the input. This analyser is still evolving, and is currently made up of about forty layers.

Such a strategy of analysis also offers the possibility to develop new layers. Besides the transcription module, which is one such layer, we seriously consider the development of a layer specifically devoted to the recognition and analysis of the "metaposition" (Ronat, 1975), which appears to be a major stylistic figure of classical verse.

### 2.2 Grapheme to phoneme conversion

The grapheme to phoneme transcription system we used follows a two-step procedure. Each pair (word, tag) is first transcribed in isolation into an abstract phonemic representation, very similar in its spirit to a underlying form. This transcription is achieved by the application of an ordered set of rewriting rules. It is enriched by two additional data. The first one is strictly lexical and flags words that are subject to diaeresis. The second one is the result of an analysis of the syntactic context to assess the degree of connection between the current word and its successor (should there be a liaison, a break, etc.?).

The second step consists in sequentially applying a series of phonological derivations, leading to a final representation of the input, which can be viewed as a kind of (quite abstract) surface form. Our system knows several such sets of phonological rules, and can produce several surface forms, corresponding to different styles of speech (formal, informal, etc.). These rules mainly follow the description of the French phonological system given in Dell (1985). A detailed presentation of this transcription system is given in Yvon (1993).

Working out a simple example should clarify the roles of the various components of the transcription module. Let’s consider the processing of the following input: "le chien mange" (the dog is eating), sketched on table 1.

<table>
<thead>
<tr>
<th>Rule</th>
<th>New representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rewrite rules</td>
<td></td>
</tr>
<tr>
<td>&quot;(le&quot;, article) -&gt; lə</td>
<td>lə</td>
</tr>
<tr>
<td>(&quot;chien&quot;, noun) -&gt; fiən</td>
<td>lə</td>
</tr>
<tr>
<td>(&quot;mange&quot;, verb) -&gt; manə</td>
<td>lə</td>
</tr>
<tr>
<td>Phonological rules</td>
<td></td>
</tr>
<tr>
<td>semi-vocalization of high vowels</td>
<td>lə</td>
</tr>
<tr>
<td>nasalization + nasal consonant delation</td>
<td>lə</td>
</tr>
<tr>
<td>deletion of final schwa</td>
<td>lə</td>
</tr>
</tbody>
</table>

**Table 1 : grapheme to phoneme conversion**

The last step of treatment is quite straightforward, and consists in identifying the syllables in the output phoneme stream.

Adapting our existing transcription system to metrometrics mainly involved the modification of this phonological module, essentially the development of a few verse specific phonological rules, which will be examined in the next section; the rest of the system remained almost unchanged.

It shall finally be noted that automating the grapheme to phoneme transcription of texts implies a certain number of arbitrary choices, from the identification of a valid phonetic alphabet, to the selection of a "correct" set of phonological derivations, and of the detailed identification of their respective scope. Our transcriptions shall be therefore regarded as abstract phonemic transcriptions.
produced for the sole purpose of counts. Neither allophonic variations, nor prosodic marks (length, strength, height) have been elicited, which distinguish our system’s output from what a "true" speech synthesis system would have produced.

The following figure summarizes the complete processing of the first verse from "La Thébaïde" by Racine.

Ils sont sortis, Olympe? ah mortelles douleurs!

Fig. 2: System overview

A last point is worth mentioning, which is not yet correctly addressed in our system: many verses are divided into two (or more) turns of speech. In these cases, the verse unity is broken, and each speech is analyzed and transcribed in isolation. This has been, as we shall explain, a major cause of errors in our syllabic counts.

3 Particularities of poetic reading

In this section, we will carefully describe how the transcription system, originally developed for standard reading, has been adapted to take into account the particularities of poetic reading.

Pronouncing verses differs from pronouncing prose in three main aspects, and this required a modification of our system both at the level of lexical word pronunciation and at the level of sentence pronunciation.

As far as the lexicon is concerned, the major difference is the diaeresis: some vocalic groups that account for one syllable in standard reading have to be pronounced as two vowels in poetic diction. For instance, the word “diamant” (diamond) should be transcribed as /dja'mɑ̃/ to reflect the standard pronunciation, and as /dya'mɑ̃/ to be consistent with verse reading. In the first case, the group “ia” will count for one metrical position, and for two in the second.

If we now consider global sentence pronunciation, most differences between the two styles of readings are easily expressed in the general framework given by Milner and Regnault (1987). As we shall see, their assumption is that the whole verse, in poetic reading, should be given the same role as the phrase in standard diction, i.e. should be regarded as a single phonological word. This radically simplifies the description of verse phonology and will be discussed when considering the special cases of mute-e and of liaison.

It should be noted that only diaeresis and mute-e elision are likely to disturb the counting of syllables, for they affect vowels. The question of liaison affects this count indirectly, for a liaison is likely to modify the status of a preceding mute-e.

3.1 Diaeresis

When two vocalic phonemes are in contact in the same word, the syllabic partition of those phonemes is called diaeresis (Mazaleyrat, 1974). As many such partitions result from the historical evolution of spelling, the analysis of diaeresis must be conducted in a historical perspective.

Originally, diaeresis was based on etymological data. The rules were simple, as stated by Grammont (1965) and there were three cases:

- The two French vowels constitute a diphthong corresponding to a single Latin vowel or a vowel and a consonant: they constitute one single syllable.
- The two French vowels correspond to two Latin vowels, that in Latin were separated by a consonant (...) or were already contiguous. In both cases, they constitute two syllables.
- Of the two French vowels, one is the final of a simple word and the other the initial of a suffix that was added to the root to produce a French derivative. Again they constitute two syllables.

Diaeresis originally caused almost no disagreement between poetic and standard word pronunciation, which simply reflected word etymology. But, as the standard pronunciation became simpler and more synthetic, the perception of increasing conflicts between both reading styles led classical metriicians to establish reading guidelines, which remained a reference during the following centuries. These guidelines took into account two main phenomena,
which affected the etymologically correct pronunciation, and disturbed syllabic counts: a general preference for synaeresis, and the institution of the so-called “three consonant rule”.

Following Mazaleyrat (1974), we can identify two major explanations for this tendency to synaeresis. First, the influence of word usage: common words (“oui” (yes), “diable” (devil)) are first concerned by this alteration, while pronunciation of learned words is more stable. Even if this accounts only for a few cases of conflicts, this linguistic pressure toward synaeresis progressively offers a new degree of freedom in poetic writing: “hier” (Yesterday) has to be counted for one syllable in Corneille’s plays (/iɛ/) when Racine conforms to the old pronunciation /iɛ/. Second, abusive analogies between phonetically similar groups also influence word pronunciations, here again in favor of the shortest form. For instance, all verbal forms ending in -ions (first person, plural) were rapidly counted as one syllable. If this was consistent with etymology in the case of subjunctive forms, which derived from Latin -eamus, it also inappropriately affected imperfect and conditional forms, which derive from Latin -eamus and should be counted as two syllables.

The “three consonant rule”, instituted during the classical era, creates new cases of diaeresis. This rule prescribes groups consisting of three consonants as valid syllable onsets. As a direct consequence, onsets composed of a stop, a liquid, and a semi-vowel were ruled out, causing the promotion of the semi-vowel into the corresponding high vowel. This rule imposes “devrions” (We should to be pronounced /dəvrjɔ̃/ or /dəvrjɔ̃/ (three syllables) rather than /dəvrjɔ̃/ (two syllables). It shall be noted that this rule is nowadays fully integrated in the French phonological system.

The system of diaeresis in classical French progressively emerged from the interaction between conflicting pressures: respect of etymological distinctions, general preference for synaeresis and the three consonant rule. A first strategy for handling this system would have been to carefully analyse the history of these conflicts. We however adopted another strategy, which was found to be more convenient for our own purpose. Rather than building etymological lexicons, we made an empirical survey of the diaeresis rules on the complete theater works of Corneille and Racine. We simply examined throughout our corpus the pronunciation of vocalic groups, and listed the forms that were subject to diaeresis. This gave way to a strictly lexical treatment of this phenomenon.

Empirical Study of diaeresis in Corneille’s and Racine’s plays.

Following Ferdinand de Gramont (1876), who had already undertaken a similar analysis of vocalic groups, we carefully examined every single sequence of contiguous vowels in the corpus of reference. This allowed us to precise the identification of morphemes or words subject to diaeresis, and led us to suggest a new set of rules accounting for synaeresis/diaeresis, based on phonetic, lexical and syntactic data, and their exceptions. The procedure and results are detailed in Beaudouin (1993).

It is likely that many writers of the classical era applied the same rules, with more or less the same set of exceptions. But some differences will most probably be found in the study of other writers, and not until exhaustive study of all alexandrines is completed shall we know the “exact” rules.

To account for diaeresis in our system, we simply introduced the lexical feature “diaeresis”. Words are marked either directly in the lexicon, or by application of heuristics. For instance, a rule states that all verbs (and all derived forms) ending in “ier” (“marier” (to marry), etc.) carry this mark. “Privilégié” (to privilege) is the only known exception, and marked as such in the lexicon. Another rule states that all proper names (with a few exceptions “Fierabras”, etc.) have to carry this mark.

This feature is then used as a meta-context to the (standard) phonological rule governing the devocalization of high-vowels: the rule is inhibited for all words carrying this feature.

3.2 The mute-E

The question of mute-e elision is quite different and cannot rely upon a mere lexical treatment. The mute-e (or schwa) refers to a phoneme which in French, can be either silent, or realized as [œ] or [œ] depending on the contexts in which it occurs. The problem of “Mute-e elision”, that is the precise listing of the contexts in which a schwa should be silent, has long been considered as a very complex question, mostly because these contexts cannot be expressed only in terms of phonological continguities, but have also to take into account the conditions of speech production, and many other sources of variability.

We will first present the views of Milner and Regnault, before eliciting our own interpretation and implementation of this theory.

3.2.1 Mute-e elision in poetic reading.

For Milner and Regnault, the same principles account for the elision of mute-e both in standard and
poetic reading, the only difference being the definition of the relevant linguistic unit in which this phenomenon is to be described: the phrase in the first case, the whole verse in the second.

These rules are quite simple and can be expressed as follows:

(1) Mute-e elision before a vowel
   A mute-e is deleted when preceding another vowel
(2) Final mute-e elision
   A mute-e is deleted at the end of the phonological word
(3) Intra word mute-e elision
   Within a phonological word, a mute-e is likely to be deleted if its deletion does not result in the creation of a group of three (or more) consonants.

Only rules (1) and (2) are mandatory, and apply both in standard and poetic diction. Reformulated in the special context of classical verse, they can be expressed as:

(1') Within the verse, all mute-e’s are not elided, except if they precede another vowel.
(2') A mute-e is not counted at the end of a verse.

Let us examine how these rules apply on the following example:

Refuse cette joie, et s’en trouve accablée
Corneille, Le Cid, verse 54

This verse contains five mute-e, (underlined in the text). The first two have to be counted, because of rule (1'). The next two shall be elided, for they are followed by another vowel. The last one is deleted by application of rule (2').

This rule formalism is indeed all we needed for a simple syllabic count. However, we decided to follow Milner and Regnault a little further and to refine this analysis of the mute-e elision. In order to account more precisely for the reality of poetic diction, they propose to distinguish among six different types of mute-e’s, reflecting both the differences of “strength” in the way these mute-e’s are read, and the underlying conflicts between linguistic and poetic structures.

The basic idea behind this classification of mute-e is that schwas that are realized as well in normal reading should be somewhat “stronger” than those which are always pronounced in verse even though they are silent or optionally silent in standard readings. As we shall see, this idea has been partially integrated into the metrometer.

3.2.2 Metrometrical treatment of mute-e

Adapting our transcription system to the poetic diction has involved the reformulation of two of the four rules devoted to the mute-e. The unchanged part concerns the status of mute-e in a vocalic context and can be stated as:

(1'') Mute-e is elided in front of a vowel
(2'') Mute-e is elided after another vowel.

The third rule is concerned with mute-e at the end of lexical words. It aims at deleting mute-es at the end of a verse, and at “weakening” as many other mute-es as possible at the end of a word.

(3'') Word final mute-e’s are weakened except in monosyllabic words, and in contexts where they cannot be elided in standard diction (“three consonants” condition, for instance). Verse final mute-e is elided.

The final rule deals with word internal mute-e’s that can be elided in normal speech. These are also weakened.

(4'') Word internal mute-e’s are “weakened” when preceded and followed by one (phonemic) consonant, except when mute-e is the first vowel, or when the following consonant precedes a semivowel.

Our system finally distinguishes three kinds of mute-e’s: those which are silent, those which are weak, and those which are strong. As this kind of distinction is sub-phonemic, we have decided to use two different symbols for the mute-e. The strong ones are noted /\ as in standard IPA coding and the weak ones shall appear as /a/.

As a result of these conventions, the following verse:

Où suis-je ? Qu'ai-je fait ? Qu'ois dois-je faire encore
Racine, Andromaque, verse 1

will be metrometrized as:


The first two mute-e belong to monosyllabic words in appearance only, for they cannot be separated from the verb, and are consequently weakened by rule (3''). The same can be said of the fourth one. The third is strong (monosyllabic word condition), and the last two are elided by application of rules (1'') and (3'').

3.3 Liaison and latent consonants

French verse being essentially syllabic, our ultimate goal was to identify syllables in the verse
3.3.1 Syllabic decomposition

The model of the syllable as implemented within the metrometer can be described as follows. A phonemic syllable is made up of an optional consonant cluster (the onset), a mandatory vowel, and an optional consonant cluster (the coda). The complete and exhaustive listing of what can or cannot be considered as a valid onset or coda is still an open question for French. In our specific corpus however, problematic words were quite rare, and we could address the question of syllabic decomposition with the following conventions.

Within a word, a single consonant between two vowels is always treated as an onset, and kept with the second vowel. Clusters of two consonants are always split into a coda and an onset, except for the clusters where the first consonant is a stop, and the second a liquid\textsuperscript{14}. Clusters of three consonants are quite rare, and in verse are always split after the first consonant. Finally, non latent word final consonants are always kept with the preceding vowel\textsuperscript{15}. These principles lead to the following decompositions:

<table>
<thead>
<tr>
<th>Word</th>
<th>Syllable Composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aricie</td>
<td>a / r i / s i</td>
</tr>
<tr>
<td>Andromaque</td>
<td>D / d r ø / m a k</td>
</tr>
<tr>
<td>phestion</td>
<td>e / f ´s / t i / ø~</td>
</tr>
</tbody>
</table>

In all cases, these conventions guarantee that a word boundary always corresponds to a syllable boundary. Not always in fact, for there is one exception with liaisons.

3.3.2 Liaison and poetry reading

Liaison in French can be briefly described as follows. When a word final consonant that is mute in an isolated word, is followed by a word starting with a vowel, it may, under certain circumstances, be promoted into a “true” consonant, and be pronounced. These consonants are referred to as latent consonants. A typical example of this phenomenon concerns the definite plural article “les” (plural of the). Pronounced / l e / in isolation, it transcribes as / l e z / when followed by a word initial vowel: “les enfants” (the children) is consequently realized as / l e z ø f ø / . The exact conditions under which this highly variable phenomenon is likely to occur have been extensively studied (see for a review \crevé (1988)).

In the specific context of poetry reading though, the problem of liaison cannot be treated as it is in standard speech, as it might indirectly have an impact on the syllabic count. The rules governing liaisons have therefore to account for the many cases where a liaison is mandatory (for metrical reasons) in the verse when it would not exist otherwise. This happens each time a mute-e is followed by a word final latent consonant, and the initial of the next word is a vowel. If there is no liaison, the mute-e will be deleted because of rule (1’’), whereas if the consonant gets promoted, the mute-e is only weakened because of rule (3’’).

Here again, our major reference is the description of Milner and Regnault. If we assume that liaison is authorized in standard reading only within a phonological word (i.e. a phrase), their assumption that the verse functions as a unique phonological word leads to a straightforward conclusion.

\begin{itemize}
\item All possible liaisons are realized within the verse\textsuperscript{16}.
\item There is no liaison between a verse and its successor.
\end{itemize}

Theoretically, these simple principles should provide a unified framework to account for both mute-e elision and latent consonant promotion.

Nevertheless, these rules are likely to promote consonants in contexts where a liaison may sound very strange for a French speaker. For example, they might lead to promote a latent consonant even if it is followed by a strong punctuation mark (a comma or a period), which is for the least unusual.

Therefore, following these authors, we have chosen to identify two different kinds of liaisons, depending on the context in which they occur:

- Liaisons that are either optional or obligatory in standard speech, and which should be "normally" realized. We shall refer to these as "D-liaisons".
- Liaisons that are not usually made, and where the pronunciation of the latent consonant should be somehow weakened. We will refer to these as "I-liaisons".

The distinction between D and I-liaisons should be marked in speech, as Milner and Regnault suggested. While D-liaisons should cause the latent consonant to be re-syllabized as the onset of the next word initial syllable, I-liaisons cause the latent consonant to be promoted in its original cluster.

As a result, all possible liaisons are marked in either manner in our system. This is why liaison is probably where our transcription system differs most from what would be pronounced on a stage. There is
indeed a general tendency in French to suppress optional liaisons in informal speech.

3.3.3 Our treatment of liaison

If the principles of Milner and Regnault are easily stated, their formalization and implementation are the much more complicated. The marking of the latent consonant is taken into account both in the lexicon, and by a simple morphological analysis that merely flags plural forms.

Applying the first principle of Milner and Regnault leads to the promotion of all latent consonants. Identifying the strength of the different liaisons is much more controversial. If an agreement seems to exist for the context where a liaison is mandatory, the exhaustive list of optional liaisons is not yet closed, and is in fact subject to variations (dependent on speakers, speech conditions, situations, etc.) and evolution.

As all we needed was a set of conventions to ground our counts, we mainly relied upon the description of liaison given by Fouché (1956), which presents the advantage of being quite exhaustive, if it does not reflect exactly current spoken French.

During the syntactic analysis, the possibility of a liaison between each pair of words is evaluated regarding the syntactic context of given words. As a result, every word separator in the input string is marked depending on whether it can accommodate liaison, or not.

During phonological processing, latent consonant truncation is performed in a way quite similar to what the "word final obstruant truncation rule" described in Shane (1969) would do. Latent consonant deletion obeys the following rules:

1. If a latent consonant precedes another consonant or an aspirated "h", it is deleted.
2. If the separator just following the latent consonant allows liaison, the latent consonant is not deleted and is re-syllabized.
3. Otherwise, the latent consonant is neither deleted, nor re-syllabized, but weakened.

Let us see in some examples how these principle affect our transcriptions. In the following verses, places where liaison might occur are underlined, and latent consonants are indexed by a small "l" in the transcriptions.

Et mes coursiers oisifs ont oublié ma voix,


Racine, Phèdre, v. 552

This verse contains three liaisons. The first (between "coursiers" and "oisifs") is possible though quite rare and is classified as an I-liaison: the latent consonant remains attached to the preceding vowel. The second (between "oisifs" and "ont") is not possible and is also an I-liaison. The last (between "ont" et "oublié") is mandatory: it is classified as a D-liaison causing the latent consonant to be syllabized with the next vowel.

When the latent consonant is an "n", an additional phenomenon has to be taken into account, namely the nasalization of the preceding vowel. As a convention, both in contexts where rules (6) or (8) apply, the vowel preceding the "n" gets nasalized. In the context of rule (7), the vowel remains unchanged. This leads for example to the following transcription:

Des crimes de Néron approuver les horreurs,


Racine, Bérénice, v. 353

There remain one case where our treatment disagrees with Milner’s and Regnault’s proposal. While they suggest that a D-liaison cannot occur at the hemistich, we have maintained this possibility, for the simple reason that we make no assumption on the kind of input we process; in particular we do not know a priori where the hemistich is located, nor even if there is one. Such cases, which imply that a phrase might cross the hemistich are quite rare, though.

Our results show that the treatment of liaison is definitely the weakest point of our system, and one of the main causes of errors. This comes on the one hand from the incomplete specification of all liaison contexts, and on the other, from mistakes about syntactic analysis. The following example, where a mistreatment of metaposition ("à la reine" should be attached to "expliquer", rather than to "vient") causes the system to diagnose a D-liaison, when there should be an I-liaison, is such a typical error.

Lorsqu'il vient à la reine expliquer son amour


Racine, Bérénice, v. 6

3.4 Evaluation of the overall system

We have evaluated our results on a randomly selected set of 116 verses in all Racine’s plays. The transcriptions produced by the system have been manually checked. On the 2949 phonemes in our corpus, 2916 (98.88 %) were judged to be correct according to our conventions.

The remaining phonemes were mainly judged as "probably not correct" (28 out of 33), and these were almost all latent consonants. Finally, a few errors (5
phonemes) were found, on which 3 were due to an incorrect assessment of mute-e strength.

Even if this test has only been run on a rather small set of verses, the mere fact that our system has been able to count twelve syllables in almost all verses of our corpora is also a positive sign of its ability to transcribe poetry into a correct phonemic representation.

### 4 Preliminary results

Our statistical analyses take advantage of the following output (see figure 3) produced by the metrometer. Since these analyses are still on-going, we will shortly illustrate the possibilities offered by our transcription system with three examples: the vocalic phoneme distribution, a quick study of mute—e distribution over the twelve syllabic positions, and an analysis of noun location in the verse.

<table>
<thead>
<tr>
<th>Verse</th>
<th>Ils sont sortis, Olympe, ah mortelles douleurs,</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonemic transcription</td>
<td>i l / s ô / s Ø r / t i z l / o / l p / â / m Ø r / t ù º / l E / d U / l</td>
</tr>
<tr>
<td>Number of syllables</td>
<td>12</td>
</tr>
<tr>
<td>Syllable location</td>
<td>1 1 0 1 0 1 1 0 0 1 0 1</td>
</tr>
<tr>
<td>Syntactic categories</td>
<td>6 1 2 4 4 0 2 2 2 0 0</td>
</tr>
</tbody>
</table>

**Reading guide**: The third metrical syllable corresponding to the phoneme sequence / s Ø r / is not located at word boundary (see line Syllable location) and belongs to an adjective (see line Syntactic categories - 2 = Adjective).

**Syllable location**

<table>
<thead>
<tr>
<th>0 = Intra-word syllable</th>
<th>1 = word final syllable</th>
</tr>
</thead>
</table>

**Syntactic categories**

<table>
<thead>
<tr>
<th>0 = Noun</th>
<th>3 = Adverb</th>
<th>7 = Preposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 = Verb</td>
<td>4 = Proper noun</td>
<td>8 = Conjunction</td>
</tr>
<tr>
<td>2 = Adjective</td>
<td>6 = Determinant and pronoun</td>
<td>9 = Relative Pronoun</td>
</tr>
</tbody>
</table>

*Fig. 3: Details of the metrometer output*
<table>
<thead>
<tr>
<th>Vocalic phonem</th>
<th>frequency</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>125282</td>
<td>14,0</td>
</tr>
<tr>
<td>θ</td>
<td>108518</td>
<td>12,1</td>
</tr>
<tr>
<td>e</td>
<td>103653</td>
<td>11,6</td>
</tr>
<tr>
<td>ε</td>
<td>97336</td>
<td>10,9</td>
</tr>
<tr>
<td>i</td>
<td>93881</td>
<td>10,5</td>
</tr>
<tr>
<td>ι</td>
<td>60123</td>
<td>6,7</td>
</tr>
<tr>
<td>u</td>
<td>59697</td>
<td>6,7</td>
</tr>
<tr>
<td>ø</td>
<td>38218</td>
<td>4,3</td>
</tr>
<tr>
<td>γ</td>
<td>36932</td>
<td>4,1</td>
</tr>
<tr>
<td>@student</td>
<td>35655</td>
<td>4,0</td>
</tr>
<tr>
<td>ιθ</td>
<td>31344</td>
<td>3,5</td>
</tr>
<tr>
<td>ε</td>
<td>26201</td>
<td>2,9</td>
</tr>
<tr>
<td>ο</td>
<td>23803</td>
<td>2,7</td>
</tr>
<tr>
<td>ο</td>
<td>17214</td>
<td>1,9</td>
</tr>
<tr>
<td>ιε</td>
<td>15630</td>
<td>1,7</td>
</tr>
<tr>
<td>ια</td>
<td>11629</td>
<td>1,3</td>
</tr>
<tr>
<td>ιε</td>
<td>10540</td>
<td>1,2</td>
</tr>
</tbody>
</table>

**Fig 4: Vocalic phonemes frequency pattern**

Figure 4, which presents the vocalic phoneme repartition, shows the high importance of mute-e in French verse. It represents 15,6 % of all vocalic phonemes. This rate is quite higher than in prose.

In the next figure, we present a first analysis of the line concerning syntactic categories: the frequency of the noun category throughout the metrical positions.

**Reading guide:** about 19 % of the verses have a mute—e as the first vowel.

**Fig. 5: Phoneme /ə/ distribution in the 12 metrical positions**

Figure 5 illustrates the well-known result that a mute-e cannot appear in sixth or twelfth positions. If this, as long as position 12 is concerned, is a direct consequence of our rules, the nearly complete absence of mute-e in sixth position (12 on 75000 verses), which has not been explicitly stated, is another sign of our system's reliability.

It seems quite obvious in figure 6 that nouns are very typical of the sixth and twelfth metrical positions: in almost half of the hemistiches here analysed, the last word is a noun. This result is a first step towards the hierarchization of metrical positions in classical poetry, which could be a dimension of our diachronic analyses.

**5 Perspectives**

The current version of the metrometer has proven it is able to provide a quite accurate representation of classical verse, including syntactic and phonemic descriptions of its structure. In this sense, we tend to consider that our first goal has been successfully achieved. This shall nevertheless be considered as a first stage in our research, and we are currently investigating in three main directions.

First, more complete statistical analyses of the corpus should allow us to extract a synthetic characterisation of Corneille’s and Racine’s alexandrine.

Second, we need to test the metrometer on other verse corpora. This will help us to extend our rules and will allow preliminary studies of the diachronic evolution of structures (Roubaud, 1978, 1988).

Finally, we wish to refine our system in order to take correctly into account typical poetic style figures, such as metaposition, and to extend our marking in order to progressively include prosodic markers (vowel length and stress marks).

At last, we shall have all the necessary information to attack the description of the all the rhythmic verse patterns, which remains one of the ultimate goals of the Abstract Rhythm Theory as proposed by P. Lusson (1973) and J. Roubaud (1991).
Acknowledgement

This paper is deeply indebted to J. Roubaud and P. Lusson who guided us all along this work.
We thank P. Constant who took the time to adapt his syntactic analyser to the verse's particularities.
We are also sincerely thankfull to F. Bimbot, J. L. Dessales, J. G. Ganascia, M. Rajman and T. Sowley for their advice and careful reading.

References


1The authors wishes to thank J. Roubaud, who put this corpus at our disposal. It should be noted that the corpus follows the rules of modern spelling, except for a few archaisms which are needed for metrical reasons.
2 Both the determiner (un) and adjective (petit) are singular, when the noun is plural.
3 The lexicon used by the metrometer has been developed at TELECOM PARIS by F. Pigamo (1990), and recognizes about 480 000 forms. It has been marginally enriched with a few exotic proper names, and a couple of archaic forms ("presques" (old almost), "jusques" (old : until), etc.). It shall be noted that this lexicon functions with a very small set (10) of syntactic categories. Additional refinements of these categories are tackled via a very rich system of syntactic features. For instance, determiners and pronouns belong to the same class, but additional features (singular or plural, etc.), allow us to distinguish among them.
4 Metaposition refers to the figure which consists in moving a complement before its predicate, be it a verb, an adjective or a noun. The following verse "Ce n'est plus une ardeur dans mes veines cachée" (Racine, Phèdre, v 305) illustrates this figure, as the complement ("dans mes veines") precedes the head of the syntactic group (here an adjective : "cachée").
5 We will call phonems the characters of the phonetic alphabet used to represent the transcription.
6 A | sign marks a weak word boundary (liaison is possible), while || marks a strong word boundary (liaison is forbidden)
7 Our transcriptions are expressed using a superset of the International Phonetic Alphabet. As we shall explain we have felt it necessary to mark latent consonants (indexed with an l) and to distinguish between two kinds of schwas ( [ə] and [əl]). In final transcriptions, syllabic boundaries appear as /
8 Grammont (1965) limits the scope of this rule to groups were the second consonant is a liquid.
9 It shall be noted that this word marking relies upon the probably incorrect assumption that within a single word, all high vowels will behave the same way. We have however not encountered in our corpus words with more than one high vowel.
10A more general hypothesis would be to consider each hemistich as an independent phonological word; this would allow the description of both the classical and epic verse in the same framework. There is, as far as we know, no single classical verse where this generalization would perturb the syllabic count. It shall be noted that this hypothesis may lead
An exhaustive description of mute-e elision in French, probably one of the most complex and controversial aspects of French phonology, could undoubtedly not be fitted into such a simple model. Nevertheless, these rules provide a simple and powerful framework in which to express our treatment of this phenomena in the special context of poetry. For a more detailed analysis of mute-e elision, refer for example to Dell (1985).

This rule is indeed more difficult to implement than it seems, for it requires the pre-evaluation of the phoneme which follows the mute-e, and which consequently forces us to take into account liaisons between words.

We do not follow here Milner and Regnault for whom the mandatory alternance between masculine and feminine rhymes should be marked in the diction. They consequently suggest to alter verse final high vowels when followed by a (silent) mute-e: the deletion of schwa is balanced by the addition of the corresponding semi-vowel. This leads for example to read “une vie” (a life) as /y n v i j/ in verse final position. We prefer to treat in the same manner all cases where a final mute-e follows another vowel, and consider balancing its deletion by the lengthening of the preceding vowel.

More precisely these clusters are: \{pbtdkgfv\} + r, and \{pbkgfv\} + l.

This is probably the most debatable convention. It is well admitted that in such cases, within a syntactic group, the final consonant can be viewed as the onset of the initial syllable of the following word (if the following word starts with a vowel), rather than as the coda as the word final syllable. Taking this phenomenon into account could be a possible refinement of our syllabic decomposition routines.

Except for a few cases, where a liaison would cause a sound clash. For example, when a liaison causes the repetition of the same consonant, etc. We have decided to ignore such refinements.

The morphological marks of plurals in French are mainly consonant final clusters, where the final consonant is latent, be it an "s", as is the most frequent case for nouns and adjectives, or one the various verbal endings ("ons", "ez", or "ent").