Daniele Fusi
University of Rome “Sapienza”

A Multilanguage, Modular Framework for Metrical Analysis: It Patterns and Theorical Issues

1. FRAMEWORK OVERVIEW & LAYERED ANALYSIS

The system illustrated here (henceforth referred to by its codename *Chiron* [1]) is built at the level of abstraction required to work with several different languages, poetic traditions, metres and texts, and as such it provides a componentized architecture. Each functionality is implemented independently, and can be replaced with others; third-party components can be developed by any user, for any specific purpose, without altering the framework. This provides all the core generic functions which would otherwise have to be repeatedly implemented, thus allowing developers of specialized components to focus on their subject. It also provides a great degree of freedom in the choice of technologies for those parts which are more likely to change rapidly following the evolution of IT, or that offer the maximum compatibility with existing systems.

At its core, the analysis flows through a chain of layers, each contributing to the shared data representing the unit of text examined. It is important to point out that the code of each layer is independent of the code of the other layers, as they must remain freely chainable in any way. Each layer operates independently, but this does not mean that it does not rely on data collected by the preceding layers. On the contrary, this is exactly the purpose of layering: each chained analyzer can use the data collected by all its ancestors, which always reside in the same shared set of segments representing the text analyzed.

[1] A previous version of this system was presented a few years ago (Fusi 2009). Since then, I have rewritten the code reusing all the relevant algorithms, theoretical background and experience from the first releases. The main reason for this refactoring is the evolution of technologies and architectural patterns, together with the requirement of a componentized and expandable architecture, open to collaboration.

For instance, the following diagram summarizes the hierarchy of the essential components for the analysis of Greek metrics. This hierarchy comprises 3 layers, named in the system lingo as *phonemizer* (phonology and prosodies), *syntaxizer* (appositives and clitics) and *metricizer* (metrical scansion):

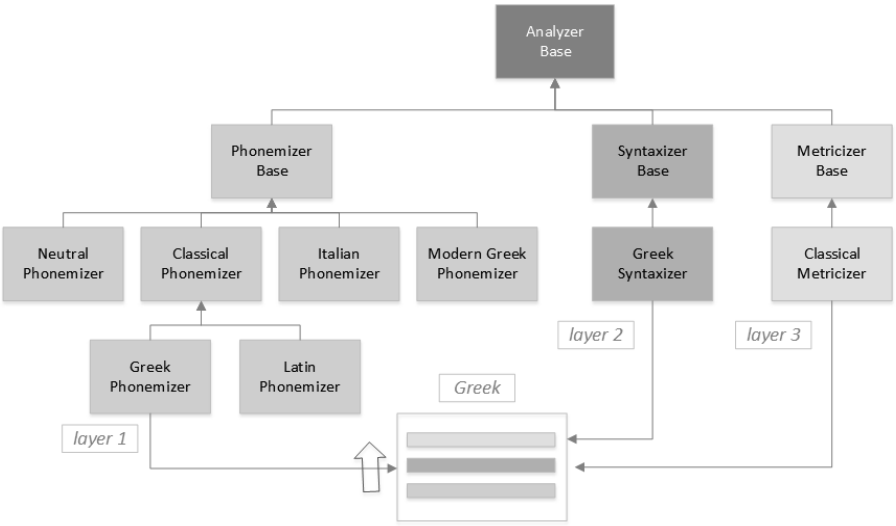


Figure 1 : Simplified diagram of Greek analyzers

As can be seen, the phonemizer specialized for Greek shares some common functionalities with the Latin phonemizer; in turn, all the phonemizers share a common functionality. The higher the level, the more abstract (and more general) the analysis. At the bottom of the chain we therefore find the Greek phonemizer; the Greek syntaxizer is located above; and at the top we find the metricizer, common to both Greek and Latin. Despite this layering, the application deals with a single analyzer component, which is just the topmost. Each query is passed to the analyzer, which internally propagates it down until the bottom is reached: there, the analysis proper starts, and proceeds upwards. This allows the application to deal with any kind of chain, from the most complex to the simplest, without any difference, as if there were only one object to interact with. Yet, each layer is specialized in its own task, thus conforming to a general separation of concerns [2].

[2] This is a well-known and widely employed principle in Information Technology. See e.g. Aksit, Tekinerdogan & Bergmans (2001).

2. PHONEMIZER

2.1. From Text to Segments

The first issue to consider when designing a system for analyzing metrics, which is essentially a regulation of selected, primarily phonological linguistic features, is how to handle written text as input, since no writing system is a faithful representation of the phonology of its language. From the standpoint of a computer system it all starts with a sequence of Unicode [3] characters. The encoding standard itself, being a true character database, provides the software with a considerable amount of information. For instance, it allows it to process any language, detect character categories, compose or decompose letters and diacritics, filter out all the irrelevant material, normalize, resolve digraphs, etc.

The input text must first be partitioned into some meaningful units. Often a unit corresponds to a line, but it can be defined at some other text range, such as a sentence, for instance to analyze prose. Whatever its type, this unit is ultimately just a sequence of characters, defined by a number of orthographical and editorial conventions. For both economic and historical reasons, such writing systems often lack the graphemes for some phonemes (hence their ambiguity); or, conversely, they have a redundant number of graphemes, or combine several characters into digraphs, etc.; in other cases, they even show graphical variants deprived of phonological meaning (e.g. the final form of sigma in Greek, or case differences). Thus, knowledge of the context is often required in order to distinguish different values of the same grapheme(s), or, in some cases (e.g. terms borrowed from another language), even a lexical resource is necessary, as single words do not comply with the standard rules of the graphical system (e.g. the digraph "c" in the Italian word *chic* = /ʃ/ against its usual value /k/ in *chicco*).

2.2. Preprocessing and Phonemization

The phonemization process leading from characters to segments starts with preprocessing, which removes all the noise and decomposes compound characters if necessary (e.g. Latin *x* into *k* and *s*, or intervocalic *i* into *ii*), etc. This does not mean that data are discarded but that they are simply removed from the text, as they may prove useful in later stages. For instance, punctuation signs are not strictly necessary when detecting phonemes and syllables, but they are important when dealing with syntax. Likewise, case differences can be useful, as in modern conventions they often indicate a proper noun, which might require special metrical treatment (the somewhat abused όνομάτων άνάγκη, especially

[3] This is the text encoding standard almost universally used nowadays, which has a strong semantic and linguistic foundation (Fusi, 2011:71-91).

relevant in less refined compositions, such as epigraphical epitaphs where the name of the deceased must fit the metre [4]).

Once this preparation has been completed, each sequence of characters is matched against the segmental sequences defined in the analysis parameters of each language, in the form of a specialized XML dialect. For instance, in Latin a single-character sequence such as *ē* is defined as:

```
<seq v="ē" seg="e">
<t n="artpt" v="fro" />
<t n="vochi" v="3" />
<t n="voc" v="1" />
<t n="voiced" v="1" />
<t n="vlen" v="2"/>
<t n="longm" v="1"/>
</seq>
```

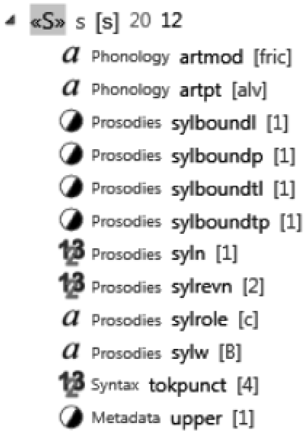
Here the sequence *ē* is described as a segmental value of *e* plus a number of traits: articulatory point (front), highness, vocoid status, voicing, vowel length, etc. Of course, there may well be cases where the grapheme(s) per se are ambiguous: e.g. Latin *i*, which can be [i] or [j] depending on the context; such cases will be marked as ambiguous and dealt with later. This enables the same preprocessing components to be shared by different languages, as at this stage we just stick to a declarative, best-guess analysis.

Preprocessing thus provides a sequence of segments, each grouping one or more characters, and temporarily assigned to their most probable phonological interpretation. Now the phonological analysis proper can take place, requiring an algorithmic approach specific for each language and context. The framework here provides a number of extension points for specific implementations: for instance, Latin *i* or Italian *i* and *u* must be analyzed in their context to determine their consonantic or vocalic role, and the same holds for the velarization of Latin *l* and *n*, or the affricate value of Italian *c* and *g*. Each segment is thus assigned a *phonological* value, represented by IPA character(s), a *text* value, which corresponds to the original character(s) the segment was inferred from, and an *alpha* value, which represents the raw, normalized letter(s) deduced from this text. This threefold value makes it possible to switch easily between the graphical and phonological representation of each segment, and provides components with access to the full data about every single segment. For instance, the first segment of the Latin *Quercus* has a text value *Qu*, an alpha value qu (lowercase), and a phonological value k^w.

[4] While avoiding mechanical explanations based on this factor, it is certain that the referential uniqueness of proper nouns makes them more difficult or impossible to replace with other terms.

2.3. Segment Extensibility: Traits

Each segment has any number of *traits*, representing any kind of information linked to it or its container. A trait is essentially a name/value pair with some metadata. Trait values can be binary (e.g. [+voiced]), numeric (e.g. an opening value), or alphanumeric (e.g. the different articulatory points: front, central, etc.). For instance, consider the first segment of the Latin word "*Sequar*":



```
◢ «S» s [s] 20 12
    a Phonology artmod [fric]
    a Phonology artpt [alv]
    ◑ Prosodies sylboundl [1]
    ◑ Prosodies sylboundp [1]
    ◑ Prosodies sylboundtl [1]
    ◑ Prosodies sylboundtp [1]
    123 Prosodies syln [1]
    123 Prosodies sylrevn [2]
    a Prosodies sylrole [c]
    a Prosodies sylw [B]
    123 Syntax tokpunct [4]
    ◑ Metadata upper [1]
```

Figure 2 : Diagnostic view of the Latin segment derived from "S"

Here the segment has a text value of *S* (uppercase, as this reflects the original text), an alpha value of *s*, and a phonological value of *s*. This segment carries a number of traits, with different data types (represented by icons) and categories, and their name/value pairs: articulatory mode and point, syllable boundaries, role and weight, etc. This structure is drawn from a linguistic analogy, but here traits have a much broader meaning, and can carry any kind of information: graphical, phonological, syntactic, metrical, etc. In a sense, the whole process of analysis is just a way of layering more and more traits on top of the same segmental base. Once this has been defined, each of the following analyzers just adds its own traits, and if necessary modifies the existing ones. Another advantage of this higher abstraction level that enhances the modularity of the system is that there is a clear separation of concerns: once the phonemizer has completed its job, there is no need to bother with complex character strings and clever parsing algorithms; we just deal with higher level data models.

2.4. Syllabification

Once text characters have been grouped into segments approximately corresponding to phonemes, the analysis proceeds to the next higher order unit, the syllable. There are a number of algorithms for the syllabification of text, used for various purposes in computer applications (hyphenation, speech synthesis, morphological analysis, etc.), each adapted to a specific use. When a language-independent algorithm is required, the most widely used implementations are probabilistic approaches, which can be either purely statistical or linguistic: examples of the former are often based on the classical “Sukhotin algorithm” (Sukhotin 1973), used to distinguish between vowels and consonants according to their distribution. This is based on two general assumptions: (i) vowels and consonants tend to alternate, and (ii) in a text vowels are more frequent than consonants. It thus operates on any corpus, starts by considering each element as a consonant, and then cycles through a number of steps, progressively detecting the most frequent element and removing it from the group of consonants. This approach can work unsupervised, and applies to any language, but it also suffers from certain limitations: first, any method based on text is affected by the orthography of the language(s), which is far from being phonological; further, such a purely distributional approach can fail, misled by the higher frequencies of some clusters. This purely statistical approach is just a method for identifying vowels, and the same applies to the detection of diphthongs, which can also be roughly defined in distributional terms (Mayer 2010).

At any rate, such algorithms define the syllabic peaks only, and we still need to distribute the consonant(s) between them. Usually this relies on typological considerations about the clusters which can occur at the beginning or end of a word (a method already applied by Greek and Latin grammarians, even against phonological likelihood, e.g. ἕ.ϰτωρ on the model of ϰτῆμα [5]), sometimes combined with principles like the *Onset Maximization Principle* (by means of which VCV is syllabified as V.CV). All these statistical approaches require large corpora to minimize their error percentage, and draw both their strengths and their limits from their unsupervised, purely distributional and/or typological approach [6].

While perfectly suitable when dealing with general-purpose automatic and unsupervised treatments, similar approaches would fall short in these kinds of specialized applications, where syllabification represents the ground for metrical analysis. In this context, a higher level of specialization is required, providing an acceptable phonological analysis from a sequence of characters, and taking into account the peculiarities of each language and tradition. Even then, this kind of automated phonological analysis is necessarily approximate, and is based on a conventional view established by scholarly tradition, mostly unaffected by more fine-grained variations in time, place, and other conditioning factors

[5] Cf. *Herodian*.2,393,33 Lentz.

[6] For instance, a Sukhotin-based approach can produce syllabifications that are clearly wrong, such as Latin *te.ne.bra.e* or Italian *qu.an.do*.

(such as morphological boundaries and speech-rate [7]). In this more linguistic and philological approach every effort is made to address each relevant language –or tradition-specific detail, so that a specialized layer is added on the universal, typological basis, and several parameters can provide different behaviors.

In *Chiron*, the generic approach to syllabification is not directly based on a text with its ambiguous orthography, but rather on the phonological analysis described above, and first of all on sonority (Goldsmith & Larson 1990), plus some typological considerations. This defines the best approximation of a syllable in purely phonological terms, which can then be refined, depending on the phonetic peculiarities of each language. The sonority of each segment leads to the definition of a curve in which each rise corresponds to a new syllable. This is the "phonological" syllable, according to the model proposed by Saussure, as in Figure 3:

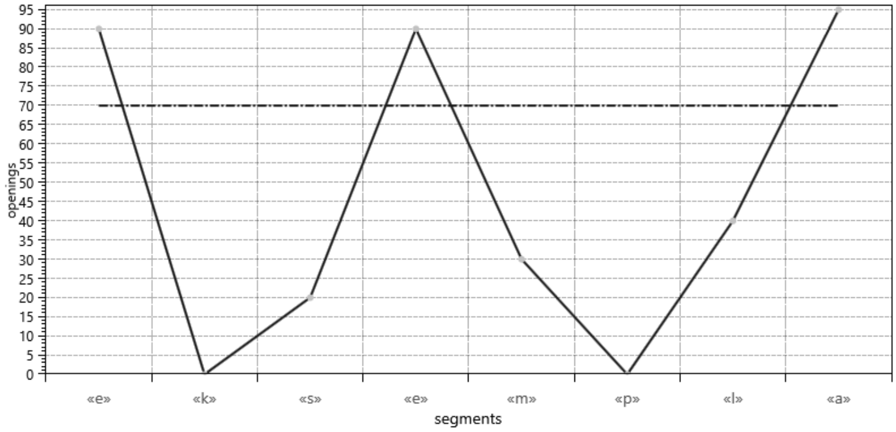


Figure 3 : Sonority curve for the Latin word exempla

However, specific languages show various degrees of departure from this model: for instance, one may conclude that Latin *stare* consists of three syllables, *s-*, *-ta-* and *-re*, just because the sonority of *s* is greater than that of the following *t* (later evolutions such as It. *istare* are a consequence of this situation). Thus, more

[7] It has been pointed out (Devine & Stephens 1994) that the fact that often in Hellenistic Greek inscriptions words are hyphenated in accordance with the principles of grammarians like Herodian –except for more "problematic" clusters like /s/ + plosive– (Threatte 1980) might well come from the habit of following an artificially slow speech rate when writing (and teaching). This is probably the case also of syllabic systems where orthographies like Myc. *ekoto* or Cypr. *timowanakotose* point to the syllabifications Ἕκτωρ and Τιμοξανα.κτος. In such cases, the metrical syllabification reflects a more "standard" speech rate, and the epigraphical one an artificially slow rate. Several other facts support the view that metrics reflected a more standard rate, like the "rhythmic *prolongement*", accentual laws dependent on syllabic weight (Wheeler & Vendryes), the morphology of the *redoublement* (with the normal pattern Ce / eCC), etc.

specialized components inherit this syllabic analysis and refine it for specific languages, marking the boundaries of “phonetic” syllables.

2.5. Probabilistic and Deterministic Approaches

The syllabification process illustrated above is an example of the more complex approach taken against usual design choices, such as that between probabilistic and deterministic systems. These approaches depend on a number of factors. When dealing with large numbers from huge corpora, and looking for general patterns from aggregated data, a probabilistic approach may prove useful and economical. A deterministic approach instead requires more complex algorithms, but in most cases it is the only way to obtain detailed data, down to the individual occurrence of each analyzed element.

Even if the general approach of *Chiron* might be described as deterministic, a distinction at the implementation level can be traced among *typological*, *specific* or *attended* analysis types, in this order of increasing specificity. For instance, syllabification occurs in two distinct steps, related to levels we can refer to with terms drawn from Saussure: the first step refers to the *langage* level, and its algorithms are based on general principles. This approach can be compared to a probabilistic one, which most times, but not always, turns out to be correct for a given *langue*. Yet, it is not based on a purely statistical collection of data from a training corpus: rather, it refers to specific syllable models. This is what we can call the *typological* level of analysis, corresponding to general principles related to the notion of *langage*. These models cannot fit any specific *langue* without some adjustments, so they are implemented by different components, one for each *langue*. Here the approach is totally deterministic, as it must best represent the peculiarities of each language. Finally, further down on the same path there is the *attended* analysis type, which requires the user’s judgment to solve ambiguities or other delicate issues. This usually happens at the *parole* level, that is, in specific passages of the analyzed text, essentially on higher level realms, such as syntax (for example, ambiguous detection of an appositive word) or metrics (e.g. ambiguous scans). In these cases, some of the *Chiron* analysis processes may pause and require user intervention [8].

3. SYNTAXIZER

A key aspect of *Chiron* is its ability to extend the analysis to several linguistic levels, insofar as they concur to solve specific issues related to metrics. In this

[8] Such a behavior is adopted for instance in disambiguating some Greek forms derived from **to-*, which can be interpreted as anaphoric pronouns (orthotonics), articles (proclitics) or relative pronouns (prepositives). Even if syntactically (and historically) the distinction among these interpretations can be tricky, on the practical side it is required. To this end, six rules are provided to achieve an automatic distinction of anaphoric and relative/article values, depending on their context and position, starting from a specific theoretical background (Monteil 1963).

context, the case of appositives (prepositives and postpositives, and clitics among them) provides a good example, as no analysis would be complete or even correct without a proper consideration of word boundaries. This is the realm of the syntaxizer layer.

3.1. The Case of Appositives

It is of paramount importance to have a practical definition of appositive words in the languages considered for analysis (in the present case, ancient Greek and Latin), especially for the nature of the metrical traditions involved, and according to the primary purposes and scope (non-lyrical versification) of this system. In what is traditionally known as *métrique verbale*, the structure of the verse is not only defined by the underlying sequence of long or short syllables, but also by the connection of words in a line. This determines the distribution of word boundaries, with their high or low frequencies. These are regarded as one of the most obvious traits of each description of metre and poetic genre: the former represent the caesural points, the latter the bridges, often described by metrical "laws". In these metrical traditions, any serious quantitative analysis about such phenomena should not ignore the fact that the notion of "word" cannot be simply identified with its (modern) graphical representation [9]. Admittedly, the issues arising from the innumerable efforts [10] to attain universal definitions of such naïve (in the linguistic meaning) notions are well-known [11], but this does not mean that we can be satisfied with a deceptively simplistic approach which just treats all the entities surrounded by spaces in a printed text as "words". Rather,

> simply by defining the phonological word in slightly more sophisticated terms than as an entity with white space on either side of it in the printed text, one can avoid the implication that the tragedians wrote hundreds of unmetrical trimesters. (Devine & Stephens, 1978:315)

In *Chiron*, word boundary types can be marked with a fine-grained distinction which takes into account more than 30 cases, combining several factors such as true/false word boundaries, hiatus, elision, synizesis and aspiration. In fact, the practical issue of appositive detection relies on the *convergence* of several criteria based on phonology, morphology, syntax, lexicon, and metrics. The

[9] With the possible partial exception of the limited case of work written "for the eye" rather than "for the ear", which appear in late antiquity as highly-learned phenomena, at times following drastic changes in the language phonology, yet following a number of patterns usually inherited from previous metrical traditions.

[10] « [...] le mot, malgré la difficulté qu'on a à le définir, est une unité qui s'impose à l'esprit, quelque chose de central dans le mécanisme de la langue ; – mais c'est là un sujet qui remplirait à lui seul un volume. » (Saussure, 1916:154)

[11] See the overview in Fruyt (1992) in a monographic issue of *Lalies*. One of the fanciest images representing this state of affairs may come from the author of one of the most famous attempts at word definition (Bloomfield, 1914:65): "It needs but little scientific reflection to make us realize that the grammarian ought by no means to extract such products [i.e. words, roots, stems, affixes, etc.] with magic suddenness, live and wriggling, out of the naïve speaker's hat". Or, in even more icastic terms, "a word is what you think is a word" (Lyons to Fruyt, 1992:**pages**).

semantic argument also retains its force, even if the progress of linguistics since the XIXth century implies less harsh statements about "meaningless" words. M. Cantilena (1995) nevertheless warns about the risk of systematically invoking words "strictly connected by meaning", arguing that this is a rather impressionistic criterion, and emphasizes the primary role of phonology: the probability of an appositive value decreases with its increasing extension, even with occasional phonostylistic variations, pointed out by A. Devine & L. Stephens (1994). A focus on phonology does not rule out other factors, such as syntax, and word order (Wackernagel's law is just the most famous of several, more fine-grained laws) can be decisive in some cases, for the appositives known as *continuatives* [12]. Some postpositives are said to become continuative when preceded by a prepositive, as e.g. in ἐκ δὲ Διός, where it is reasonable to assume that the preposition (and proclitic) ἐκ creates the expectation of a word to lean onto, which cannot be obliterated by the small postpositive (and enclitic) δὲ just before Διός. Finally, as for metrics, the risk of circularity can be avoided by the dynamic nature of such systems, which can repeat their analysis by adjusting their parameters according to the results. This set of criteria might thus be enough to provide at least a working hypothesis for defining appositive words, taking into account what M.-J. Reichler-Béguelin (1989) calls "prototypying", i.e. instead of trying to define well-cut categories, when dealing with naïve notions such as *word*, we should consider that not all the members of this category are representative of it in equal measure. Rather, they are located on a "typicality gradient", ranging from the prototypical members which bear all the traits of the category, to the peripheral members, which just possess part of them: in the category "birds", for example, a canary is commonly considered a more "typical" member than a penguin.

3.2. Practical Impact on Data

It is crucial to take into adequate account the appositive status of the words constructing the verse. Here I will limit myself to providing evidence for the relevance of appositives drawn from data analyzed using *Chiron*, by summarizing it in a few charts or verses.

First of all, as we are interested in demonstrating the impact of appositives in the numerical data collected about word boundaries, we should emphasize their distribution in the way which best isolates this phenomenon from others affecting it in the verse instance, namely the distribution of different quantitative patterns. The two phenomena, which could not be distinguished when dealing with the real language as constrained into metrical shape, have long been studied especially for the Greek hexameter. One of the best examples of this connection is

[12] "Usually an enclitic cannot absorb the force of a preceding prepositive, but can act only as a bridge" (Bulloch, 1970:262). Other traces of this belief appeared much earlier, even if as cursory remarks, as in Wifstrand (Wifstrand, 1933:74; 41): "ein postpositivum macht nicht nur die vorhergehende, sondern auch die nachfolgende Wortgrenze unklar".

provided by O’Neill (1942), who completely ignored the question of appositives and caesurae, shifting his attention to what he called the “localization of words”, i.e. the distribution of words with different quantitative patterns in the verse. Even if the theory behind this paper is nowadays unacceptable, it is useful to note how there the study of caesurae was entirely based on observation of the frequency and position of the sequence of long or short syllables defined by spaces. This implies that there are at least two different factors at play: (a) the distribution of quantitative patterns, and (b) the distribution of word boundaries. As for the hexameter, in traditional terms (which can be adopted here for the sake of the brevity and simplicity of the present discussion) point (a) equates the frequency and distribution of dactylic and spondaic “feet”, while (b) corresponds to the frequency or rarity of word boundaries (caesurae and bridges). Thus, as our aim here is to show how appositives affect the statistics related to word boundaries, we may wish to try to keep it as isolated as possible from other, concurrent factors defining the structure of the verse, such as the distribution of dactylic and spondaic feet. In other words, here we are emphasizing the distribution of word boundaries, and we do not want to be misled by the fact that some of the possible positions in the line are biased by the dactylic or spondaic realization of the foot and their change over time. Figure 4 summarizes this fact by showing the distribution of spondaic feet in the Greek epic hexameter from Homer to Colluthus:

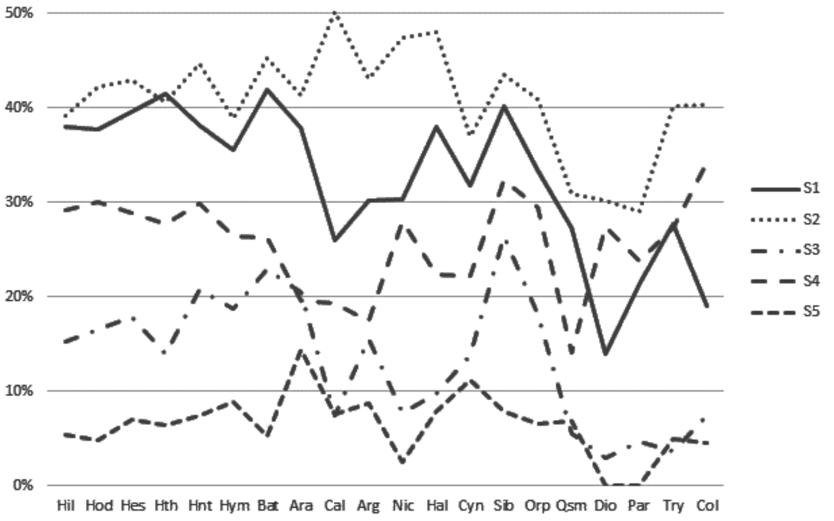


Figure 4 : A bird’s eye view of the distribution of spondaic “feet” in the Greek epic Hexameter from Homer to Colluthus (about 90,000 lines)

To properly isolate (even if artificially) the distribution of word boundaries from that of quantitative patterns, given that we can exploit the full details about each datum collected by the system, we can just calculate the frequency of word boundaries relative to the frequency of each single pattern, rather than of the total count of the lines. The following chart presents such data relative to one of the works analyzed in the corpus, Aratus' *Phaenomena*. The true word boundary distribution is the highest solid line, and it is easy to see that it follows a curve which reflects the traditional positions for caesurae (in its peaks) and bridges (in its valleys). The false word boundary distribution is the other solid line, while the dotted areas represent their "weighted" versions, i.e. their frequencies calculated on the compatible patterns only, rather than on the total number of lines (=1,153).

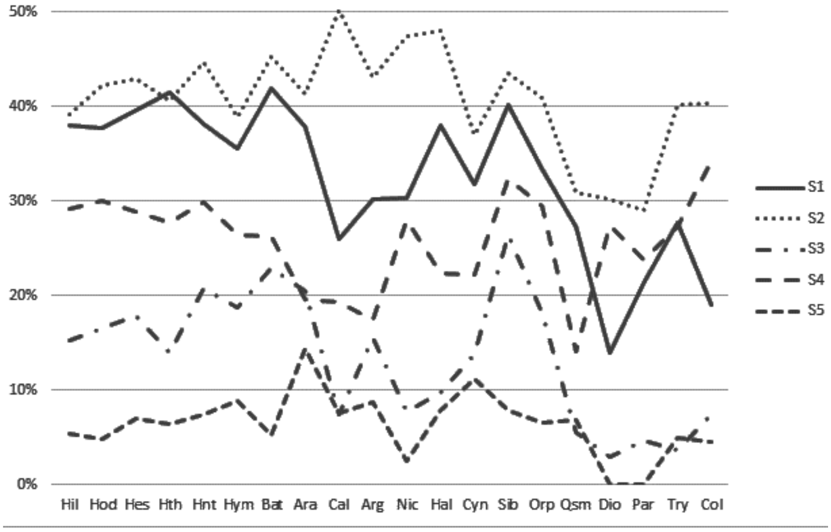


Figure 5 : Distribution of word boundaries in Aratus
t=true, f=false, tw=true and 'weighted', fw=false and 'weighted'

There is a clear difference between "weighted" and non-weighted frequencies: the former are always higher or at least equal to the latter, and therefore appear more salient, especially wherever the distribution of dactyls and spondees is markedly different: see e.g. the 1st, 3rd and 4th feet and compare with the following chart, showing the frequency of dactyls and spondees in the same text:

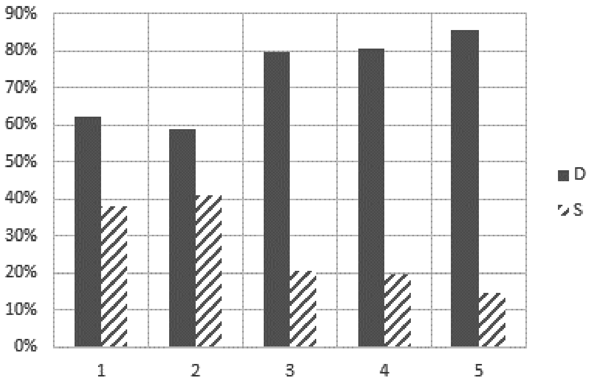


Figure 6 : Distribution of dactylic (D) and spondaic (S) “feet” in Aratus

Also, it is easy to grasp the impact of a proper treatment of appositives: this is already clear in the curves for true and false word boundaries, which, especially on some key points (see e.g. the 3rd foot), show a noticeable shift towards bridge positions. It can be seen even more clearly in the following chart, which shows the curves of the “weighted” frequencies against their sum:

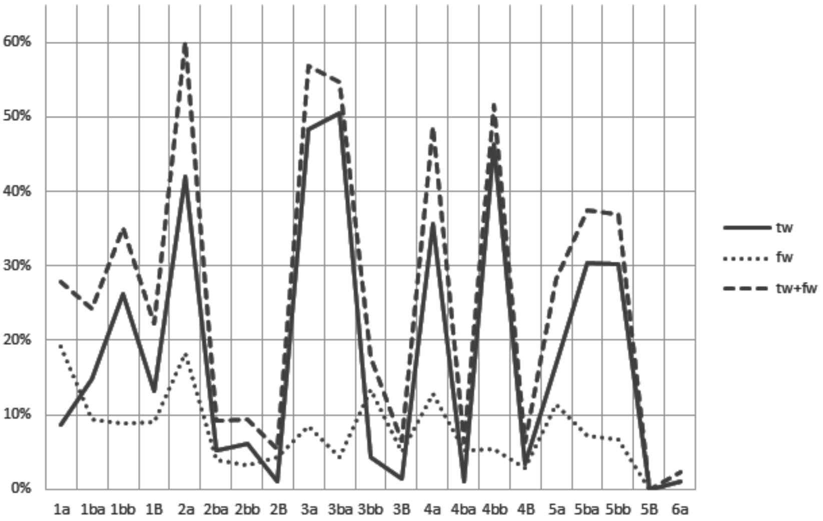


Figure 7 : Distribution of true, false and true + false word boundaries, all “weighted”, in Aratus

Here, the sum represents what would be considered the frequency of word boundaries without distinguishing between true and false, and the impact of appositives is strong. For instance, in the sum curve the relation between the feminine and masculine caesurae appears to be inverted (with the masculine prevailing, while it is well known that the feminine increasingly prevailed from the Hellenistic up to the Imperial age). Moreover, the values in the second foot are clearly biased, as the word boundary peak appears to be even higher than that of the main caesura of the verse in the third foot. The true-word-boundary curve instead reflects the expected scenario, which can be confirmed not only by the literature on the subject, but also by comparing the charts for all the other texts analyzed in this corpus and their emerging tendencies.

This example should be enough to show that taking appositives into consideration is not an option, especially in view of the fact that here each single datum can be reused and freely combined with others, for any desired type of research (see the above use of the distribution of dactyls and spondees to calculate word boundary frequencies), so that transformations based on incorrect data would be completely misleading.

While aggregated data best show the huge impact of appositives, we can also introduce some concrete examples. We will limit our examples to a relatively short text that conforms strictly to most of the hexameter laws, see e.g. *Aratus' Phaenomena*. If appositives are not taken into account, a line like 499 would break both Hermann's law (avoidance of word boundary after the two syllables of the 4th dactyl) and Meyer's third law (avoidance of word boundaries both after the 3rd and 5th *longum*):

^{1}τὰ τρία ^{2}δ' ἐν περά^{3}τῃ θέρε^{4}ος δέ οἱ ^{5}ἐν τροπαί ^{6}εἰσιν.

Here we would have to count a word boundary after the fourth "trochee" between δέ and οἱ, which is totally unrealistic even from the simplest phonological standpoint (both are monosyllabic enclitics), thus violating Hermann's law. Further, the two word boundaries after περάτῃ and ἐν would break Meyer's third law, which is also ruled out by the fact that there can be no true word boundary after a proclitic. Lastly, the two word boundaries after Θέρεος and ἐν (again, a proclitic) would break Tiedke's law (avoidance of word boundary after both the 4th and 5th *longum*). All these are consequences of the fact that in this line we find 11 "graphical" words, but only 4 "metrical" words.

The same scenario is found in line 716 ^{1}Ἡνίο^{2}χος φέρε^{3}ται μοί^{4}ρῃ γε μὲ^{5}ν οὐκ ἐπὶ ^{6}ταύτῃ; or in line 1097 ^{1}ἀμη^{2}τῷ, μή ^{3}οἱ κενε^{4}ὸς καὶ ἀ^{5}χύρμιος ^{6}ἔλθῃ which would break both Hermann's and Hilberg's (avoidance of word boundary after the 2nd "spondee") laws; etc. Such examples, which are numerous, appear more salient due to the fact that they refer to "bridges", which are usually easier to spot using a traditional approach. However, if we move towards the "positive" side of the matter, following the view introduced by H. Fränkel (1955), and look at the frequency and distribution of word boundaries and the structural patterns emerging from them, the scenario is even worse: word boundaries seem to be much more numerous than they really are, thus obfuscating the paramount relevance of caesural points in the hexameter. Looking at charts like 4 above, one might even question the well-known fact that the feminine caesura progressively prevailed over the masculine in the history of this Greek metre; and there would be a great number of verses where one might wonder whether to "choose" a masculine or feminine caesura, just because spaces are printed at both locations, as in *Arat*. 1097 quoted above. Even without taking the metrical tradition into account, the caesural status of some word boundaries against others should simply emerge from the observation of their frequency, as patterns are established by repetition, as perceived first of all by a public of listeners (what Rossi (1963) used to call "metrica aurale"); but failing to correctly distinguish between false and true word boundaries would totally scramble our data.

Thus, the main task of the syntaxizer is to provide a reliable detection of clitics, appositives, and "lexical" words. Like any layer, it uses the data collected by its ancestors. First of all, the segmentation done by the phonemizer is essential for the syntaxizer, which just deals with tokens ("words"), rather than with single segments; its main job consists in matching each token against the list of appositives, thus detecting clitics and appositives. This requires a somewhat complex algorithmic approach, as words in the line are affected by many sorts of syntagmatic modifications, and additionally some of them can be ambiguous (e.g. οἱ might be either a proclitic or the enclitic pronoun from **sw*). Moreover, a correct analysis involves concepts which are intrinsically syntagmatic, such as the notion of *continuatives*. The lower-level entities delimited by the phonemizer are also useful for the syntaxizer itself, as syllables are required, e.g. in order to correctly detect some cases of syntagmatic modifications in clitics [13]. The same holds for continuatives, which not only intrinsically require a syntagmatic definition, but also involve the count of connected syllables [14].

3.3. Metricizer and Interactions among Layers

The main task of the next level in the hierarchy, the metricizer layer, is to provide the metrical scan of each verse, thus requiring all the phonological data. Instead,

[13] A sequence like λόγου τινός shows an enclisis accent on the second word, which might prevent it from being detected if we just looked up the normalized token, as the appositives list contains the unaccented form τινος. The syntaxizer is smart enough to know that a bisyllabic enclitic following a paroxytonon gets an enclisis accent; thus, if a bisyllabic oxytonon is not found, it can further be looked up in the same list after removing its accent.

[14] Another example of revision of criteria based on data outcome (as for μάλα below) is provided by the emerging need to set a threshold to the extent of the resulting appositive sequence. We counted, for instance, several violations of Varro's and Lehrs' laws, which are among the strongest tendencies in the epic hexameter, e.g., *Il.* 12,132 ἤριπε δ᾽ ὠς ὃτε τις δρῦς ἤριπεν ἤ ἀχερωῒς or *Hes. Th.* 291 ἤματι τῷ, ὅτε περ βοῦς ἤλασεν εὐρυμετώπους: in similar cases, if τις and περ were treated as continuatives, this would not only produce a verse without a caesura in the third foot (against Varro's law), but also with a word end right after it (against Lehr's law). If instead we set a limit to the extension of continuative groups, tentatively equal to 3 syllables (according to other well-known facts in Greek phonology, mainly connected to accentuation), all such issues can be resolved.

we could even scan a line without the data collected by the middle syntaxizer layer. Of course the scan itself would provide only a very incomplete metrical analysis, insofar as it just shows the way in which quantitative constraints are met; but this is far from providing a full account of how words are laid out in the concrete verse instance with its structural building blocks *(cola)*. Once the analysis has been completed, it will be up to the components' named observers to extract any specific information related to prosodies, word boundaries and bridges, metrical laws and patterns, etc. All this requires detecting appositives, even when not directly dealing with word boundaries, because we cannot know in advance how the observed data would be combined by the end user. Conversely, the metricizer may modify prosodic data. For instance, it might be the case that a predefined choice for the syllabification of a *muta cum liquida* group happens not to be the correct one for that specific verse, which otherwise would not scan. In this case, the metricizer moves the syllabic boundary (e.g., Homeric Ἀ.φρο.δί.της = **??**– against the expected heterosyllabic treatment, because this would be incompatible with the metre). The metricizer can also add phonological data by inferring vocalic lengths from the metrical context, potentially improving the prosodic thesaurus. Similarly, the syntaxizer too may alter some phonological data, for instance, those related to accentuation, which are better defined in the light of appositive detection.

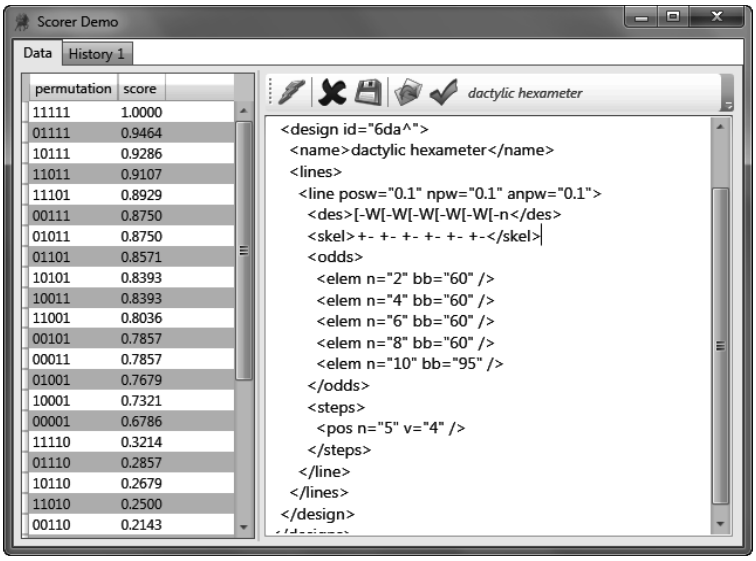
As the metrical analysis is usually the topmost layer and is closely connected with language and poetic tradition, it represents the most variable component in the system. In fact, it is easy to understand that the metrical scansion of a Greek or Latin verse (starting from a sequence in which each syllable with its weight must be fitted to any of the possible implementations of any verse design) is a different process from the scansion of an Italian verse (which basically just relies on a syllable count and accentuations). The framework is, however, abstract and modular enough to deal with such differences. The reference to some verse design is instead common to all metricizers, and uses a component which applies to any design, whatever its metrical tradition. Each metre is fully described in external resources, by a proper XML dialect. Typically, a Classical verse design implies a number of possible verse instances, defined by combining all the possible syllabic implementations for each element. For instance, even a rather simple design such as the dactylic hexameter comprises 32 possible instances, generated by combining the double implementation (either as a long syllable or as two shorts) of each of its 5 *biceps*. The metricizer must thus be able to generate all these permutations, and select the one which best fits the sequence of syllables being analyzed. Of course, not all the permutations have the same frequency. The details may vary for each design, literary genre, style and period, but users can define a set of parameters to help the metricizer generate the most probable permutations first. The parameters are based on a number of general principles for Greek and Latin metrics, which are used to calculate a "score" for each permutation. The higher the score, the higher its expected frequency. Thus, each verse design also includes some general data used to calculate a score when defining its implementations.

This is only a best-guess optimization, but it can help especially when dealing with highly variable designs. For Greek and Latin, the parameters rest on some common-sense considerations:

- each element in the design shows a preferred implementation, so that the higher the preferred implementations count, the higher the frequency of the instance (within a limit imposed by the avoidance of monotony, even if this is subject to radical changes in time);
- the preferred element implementations tend to occur more frequently towards the end of the line. For instance, it is well-known that the dactylic hexameter (5th dactylic foot) is more frequent than the spondaic one (5th spondaic foot). Thus, the preference for a specific implementation must be combined with its position, which usually increases non-linearly;
- several dispreferred implementations tend not to occur in a consecutive sequence. For instance, two consecutive spondees in a hexameter are less common than two spondees interrupted by a dactyl.

For instance, the first permutations generated by this scoring system for the epic Greek hexameter are (using D to represent a dactylic foot and S for a spondaic foot): 1) DDDDD (all the preferred implementations of each *biceps*), 2) SDDDD (1 dispreferred implementation in its less relevant position, i.e. the farthest from the end of the line), 3) DSDDD, 4) DDSDD, 5) DDDSD, 6) SSDDD (and not DDDDS, because the weight assigned to the 5th *biceps* is greater than the one assigned to a dispreferred implementation), etc.

Adjusting the weight of all these parameters in each design to obtain the best scoring can be tricky. To this end, the system offers a number of tools for interactively testing each design, evaluating the differences introduced by each change. The screenshot below shows the user interface for tuning a verse design with relation to its score: users can edit the design and generate all the permutations, together with their score.



permutation	score
11111	1.0000
01111	0.9464
10111	0.9286
11011	0.9107
11101	0.8929
00111	0.8750
01011	0.8750
01101	0.8571
10101	0.8393
10011	0.8393
11001	0.8036
00101	0.7857
00011	0.7857
01001	0.7679
10001	0.7321
00001	0.6786
11110	0.3214
01110	0.2857
10110	0.2679
11010	0.2500
00110	0.2143

```
<design id="6da^">
  <name>dactylic hexameter</name>
  <lines>
    <line posw="0.1" npw="0.1" anpw="0.1">
      <des>[-W[-W[-W[-W[-W[-n</des>
      <skel>+- +- +- +- +- +-</skel>
      <odds>
        <elem n="2" bb="60" />
        <elem n="4" bb="60" />
        <elem n="6" bb="60" />
        <elem n="8" bb="60" />
        <elem n="10" bb="95" />
      </odds>
      <steps>
        <pos n="5" v="4" />
      </steps>
    </line>
  </lines>
</design>
```

Figure 8 : User interface for tuning a verse design in relation to its score

Changing even a single parameter may lead to a fully reordered set of permutations. To appreciate the effects of such changes, this tool offers a historic view where each series of permutations is laid next to the other, and each item of each permutation is linked to the same item in the previous one. The link is graphically represented by a line, red when the item has moved down, and green when it has moved up:

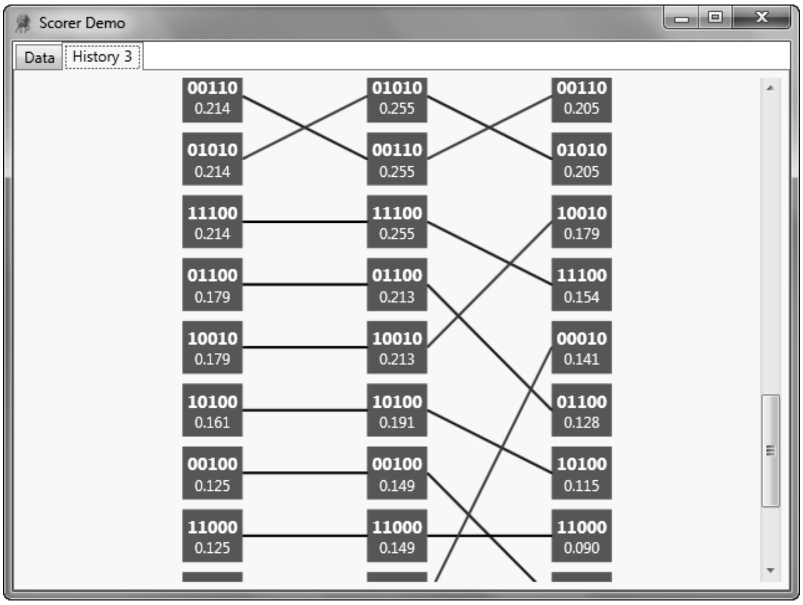


Figure 9 : Comparing the effects of score parameters

As several elements in a verse design (such as *anceps*, *biceps*, etc.) usually allow different syllabic implementations, the job of a Greek and Latin metricizer is to determine the right verse instance, by comparing the abstract verse design with the concrete text being scanned. On the design side, the metricizer can generate all the possible instances from a design. On the text side, the metricizer tries to match each compatible instance, altering its prosodies where required, and finally selecting the right scan. At this stage, it can also make use of its vocalic lengths thesaurus to automatically add new lengths, deduced from matching the instance [15]. This configures a system which is able to "learn" from its analysis experience: the more verses it scans, the more vocalic lengths it can deduce. Once added to the prosodic thesaurus, these lengths become available earlier, to the phonemization process itself, which can directly inject them during its analysis, thus removing more and more uncertainties about prosodies [16].

[15] For example, in a hexameter line beginning with μῆνιν ἄειδε the length of η is known from its grapheme, the length of ι from the word accentuation, while the length of α can be deduced from its metrical context.

[16] This is especially useful in a language such as Latin, where no letter implies the notation of vocalic length, or, in ancient terms, all the vowels are *anceps* (Rossi 1963).

4. DATA COLLECTION *VS* INTERPRETATION

4.1. Observation Scope

The main aspect of quantitative analysis in metrics is that you may be interested in a set of data which is usually much larger than the scope of your specific investigation, as metrics builds patterns from language features, which in turn are dependent on each other. Moreover, when dealing with historical developments, features not relevant for a metrical tradition may acquire importance in a changed linguistic environment. The latter is often the case of accents when studying the transition between Classical and Medieval metrics, both for Greek and Latin languages. For instance, in Classical (quantitative) Latin metrics, accents play no relevant role; still, they may show a tendency towards some patterns, just because these are the combined effect of the regulation of syllabic quantity, word length and order, verse design and word boundary distribution. Yet, when quantities begin to fade in the perception of speakers, it is precisely the distribution of accents and syllables which becomes the essential linguistic feature in building a new metrical tradition.

Nowadays, the complexity involved in the rise of a new poetic tradition, such as Medieval Latin poetry, is well-known thanks to D. Norberg (1958), who argued against the more simplistic view that early Latin rhythmical versification limited the regulation of accents only before a verse clause. Rather, there are a number of different regulations based on several ways of imitating what was felt as the rhythm of Classical versification: in this kind of poetry it is the word accent of Latin prose (and not the so-called metrical *ictus*) which provides the basis for the rhythm. Yet, this happens in the context of models which were regulated by different principles, first of all syllabic quantity, As Latin accent depended on quantity, the complex interplay between the distribution of quantities, word boundaries and accents often determined some accentual patterns. Even if these patterns, at least in Classical Latin, were just the effect of the regulation of other linguistic features, they appeared to be the truly audible and reproducible criterion for reinterpreting the model in the new linguistic context. Of course, not all the imitations are equal: some poets strive to fully imitate the accentual patterns; others limit themselves to the end part of the line, while yet others try to imitate different (connected) features such as syllable or word [17] count. It may even happen that in such imitations of quantitative verses through other linguistic features some poets accidentally manage to create what seems to us a regular quantitative verse: but here the linguistic features being regulated by this new metrical tradition are completely different.

[17] Incidentally, referring to the above hint at the different relevance of clitics, appositives and fully "lexical" words, it is useful to point out that in this case (imitation of word count) only "true" words are considered; monosyllabic prepositions, *et* and *ut*, are not counted (Norberg, 1958:131-132).

An analysis framework like *Chiron*, which can record every detail of each single phoneme, syllable, word and line examined, storing them in a corpus which can later answer any kind of queries about the collected data and their interactions, provides a statistical ground for researching these transitions between different, yet connected poetic systems, or the changes which affect the history of the same tradition in the course of time. I have attempted to provide a methodological example of both cases in two studies based on previous versions of the system discussed here, for both Latin (Fusi 2002) and Greek (Fusi 2004) poetry. In the former study, I examined the poetry of the late Latin poet Luxorius, showing that the patterns emerging from the data seem to point to a set of typological deviations from the Classical quantitative poetry being imitated, rather than a chaotic bunch of unrelated and even contradictory errors. In the second study, I proposed a more historically-oriented and data-aware interpretation of two famous Greek hexameter laws, Hermann's and Lehrs' law. The data show that the latter is at least initially a syntagmatic rule (i.e. a rule which applies only in specific contexts), while Hermann's law can be partitioned into two subdomains, of which only one can be considered as relevant (i.e. an effect explicitly sought by the poet). The other is simply the effect of a combination of factors (one linguistic, the rarity of "lexical" monosyllabic words, and the other metrical, the frequency of word boundaries in specific positions). Later, even if the evolution of the Epic hexameter shows a tendency towards the reduction in the frequency of these conditioning factors, violations of Hermann's law do not increase. My explanation for this phenomenon is tied to the rules of the literary genre: what first was (at least partially) a byproduct of the combination of conditioning factors later became an effect explicitly sought by poets, even when these factors decrease. The original phenomenon had become an unwritten rule established by the authority of their venerable literary models.

Given this variety in the consumption of data produced by analyzers, it is highly convenient to keep the analysis itself separated from the observation of its results. This allows the full analysis to be implemented independently, while providing any number of specialized data collections depending on the specific aim of each research investigation.

4.2. Observing Data

Once analyzers have stored their data in some kind of repository, it is up to the researchers to select the subset they are interested in, climbing up the hierarchy from the segment with its traits to syllables, tokens, phrases, sentences and works. The central component here is a *gatherer*, i.e. a component which grabs analysis data from a metrics repository and collects several observations, storing them in a metrics corpus. A gatherer contains any number of pluggable components, the *observers*, each specialized for a specific phenomenon. Observers are designed in such a way that they refer to the smallest useful observable unit, so that their results can later be queried by combining them into a more meaningful and customized set of data. For instance, the user may have an observer which looks only at hiatuses, another only at elisions, another only at word boundaries, and another only at the elements solutions, etc. Each of them stores its observations in a shared repository, and the user can later query it by asking complex questions such as “show me all the cases of hiatus without elision at a true word boundary after the third foot implemented by a resolved *biceps*”. Thus, the system provides a sort of true virtual metrics laboratory, where each user is free to experiment by combining observations and testing hypotheses against the results. There are about predefined 30 observers, either general or specific to well-defined languages or verse designs, so that every possible relevant aspect of the collected data is covered.

4.3. Refactoring Analysis Methods from Data

Observing the collected data often provides strong hints for refining the analysis methods according to the results. In this scenario the algorithms and resources behind the analysis, which can be very complex, can be adjusted according to the outcome of a first analysis pass. This is the essence of any unbiased data-driven metrical research (Maas’ (1962) well-known *observatio*), where scholars observe data without any preconceptions, and start formulating hypotheses to explain the emerging patterns, or to question the data themselves, i.e. the method of their collection.

A typical case occurs when the general picture emerging from the data appears clear enough to define strong tendencies, and yet a set of (relatively) relevant exceptions occur. In this case it is necessary to shift from the observation of aggregated data to their unaggregated details, in search of patterns which might provide hints for adjusting the method of analysis. For a simple case, take the role of intensifiers such as μάλα: initially, I tentatively classified this as a postpositive, following the list provided by Vendryes (1945:107), but this appeared to conflict with the strong usage of this word as an intensifier (as Vendryes himself suggests): thus, a tentative phonological criterion contrasted with a strong syntactic one. According to the theoretical principles sketched above, it was safer to provisionally classify μάλα as an enclitic, and let the data refute or confirm this. The analysis evidenced very strong tendencies for many of the traditional hexameter “laws”, among which Lehrs’ law (avoidance of word boundary after the 3rd foot when there is a word boundary in it), Varro’s law (i.e. the requirement of the “main” caesura at the 3rd foot) and the avoidance of final monosyllables. On examining all the verses listed as violations to Lehrs’ law, it became clear that about 90 of them could be traced back to cases involving μάλα, as e.g., *Il.* 10,289 κεῖσ΄ ατὰρ ἄψ άπιὼν μάλα μέρμερα μήσατο ἔργα, *Od.* 15,556 ἔνθα οὶ ἦσαν ὕες μάλα μυρίαι, ἧσι συβώτης, *A.R.* 3,1368 τόν ΄ρ΄ άνὰ ΄ρεῖα λαβών, μάλα τηλόθεν ἔμβαλε μέσσοις, *Nonn. D.* οὕτω σῶν βλεφάρων μάλα τηλόθι καὶ σὺ τινάξας, *Q.S.* 4,424 ὑδρηλῆς καπέτοιο μάλ΄ άγχόθι τηλεθάοντα, etc. Moreover, *Arat.* 907 δειδέχθαι άνέμοιο νότου βορέω δὲ μάλα χρὴ would also break the avoidance of the final monosyllable, while hundreds of lines would break Varro’s law, e.g., *Il.* 5,471 ἔνθ᾽ αὖ Σαρπηδὼν μάλα νείκεσεν Ἕκτορα δῖον. All this provides convincing evidence for the (at least aprioristic) prepositive nature of μάλα in that corpus.

This is a simple example of how metrical criteria deduced directly from the observation of analysis results can contribute to redefining the algorithms. Obviously, it is up to the scholar to decide on the relevance of this criterion, but once the system parameters have been adjusted, it is easy to repeat the full analysis and test the new hypotheses against the new results. This provides a sort of regression test, which can always be run on the whole set of data until the system appears to be fine-tuned.

4.4. Metrics Corpus

The *metrics corpus* is the set of observations produced by gatherers. Rather than a closed set of data, it is a truly interactive and composable research tool, which any user can continuously reshape, enrich, and query from any desired perspective. What makes this possible is the modular architecture of the system, with its clear separation of concerns. Once the analysis is complete, the output is represented by a set of segments with their traits, lending themselves to any desired observation. It is only at this stage that interpretation proper comes into play, in the form of observers that select from these segments all the data relevant for their job, process them, and output their observations in the metrics corpus, which allows for both aggregated and detailed queries.

For instance, think of a complex problem such as the analysis of the potential interactions between versification and accent distribution in Greek over time. First comes the basic analysis performed by the system framework, which should not involve any specific bias with regards to its task, and only later comes interpretation(s), which will vary depending on the purpose of each specific study. The first step should just collect the raw distribution of accents. Once these data have been stored, their interpretation comes into play; each observer submits a set of queries to obtain the data required to test a working hypothesis, or to collect some specific combinations and look for emerging patterns. The user may wish to plug in a component which treats all the different accent types as equal, when dealing with late antique or Byzantine poetry, or rather focus on the transitional period between two different versifications (quantitative and accentual), testing whether patterns emerge which require special consideration: for instance, what W. Allen (1973, 1987) defines as a *contonation*, graphically corresponding to the circumflex (monosyllabic contonation) or acute (bisyllabic contonation) accents. It is worth emphasizing that these data are not directly drawn from the input text; rather, they are drawn from the output of the first stage of fully detailed analysis, ending up with a set of objects, rather than with raw figures. In other words, the first stage provides data that represent the unique content of the analysis, and a second stage concerns their consumption, i.e. the variety of ways in which they can be regarded and interpreted, depending on different scholarly scenarios, in a well-known IT paradigm: one content, several presentations; one system, several languages, metrical traditions, metres, levels, observations, and extensibility points.

A final, yet key remark remains to be made, especially for those readers who look with too much, or too little, confidence to computer analysis: as with any computer-based system, it is always a matter of compromise and balance. We will probably never have a computer system capable of *automatically* taking into proper account any single marginal case one can think of. In some cases, there may even be considerations deriving from aesthetics, emphasis, semantics, etc., that deserve a special or unique treatment. However, this does not imply that our efforts in building such systems are useless, just because they are not capable of all the subtle distinctions made by human scholars. First of all, this is not their job. Such systems are *not* intended to replace human scholars; they are only tools, which provide scholars with data which it would be practically impossible to obtain manually. Their job is to help where humans fall short, i.e. dealing with huge amounts of data, whatever their level of detail. In such corpora, the percentage of marginal cases is mostly statistically irrelevant: thus, when looking at aggregated data, marginal cases cannot bias the results. The job of the system is to try to automatically provide the best account of data it is capable of, within acceptable limits of fairness, and using well-defined criteria, so that the analysis can be not only implemented, but also reproduced, and thus verified, by other scholars. Often, this will force us to make difficult choices, at least in the context of the corpus we are analyzing. It would be easy here to question some of them, by pointing to conflicting samples, but this would fail to take into proper account the nature of similar tools. Rather, what is relevant here is that these samples form a minority group. Further, and more importantly, this does not rule out the possibility of taking even such conflicting cases into adequate account: in similar systems the analysis proceeds by progressively refining its results, flowing through several layers. At a first stage it may well happen that a "best-guess" choice is applied; yet later this can be fixed. This happens, for example, in the case of prosodies, as explained above. This approach is just a way of solving complex problems by splitting them into smaller pieces, following design patterns which fit the architecture of modern IT systems, which have to be highly modular and independent. The nature of the fix can vary: it can be automatic, as in the case of prosodies; or it can be provided by human intervention. I have already pointed out that this is one of the factors taken into account by the system, at several different levels, either before or during the analysis. The system does not only provide aggregated data: on demand, it can provide any detail by listing each relevant verse. This allows scholars to examine the results line by line, detect potential problems, and then either (a) redesign the algorithms or their parameters and repeat, or (b) mark some marginal cases as such, so that the system will treat them as desired. Of course, the choice between (a) and (b) depends on the number and nature of the issues arising from the automatic analysis. If some of them can be traced back to a common reason, and this has support from theory, we will choose (a); if instead there are very special reasons to treat some specific samples as marginal cases, without motivating a change in the system, we will choose (b), and add manual markup to the input text, to force the analysis in the intended way. Ultimately, this kind of analysis is never a single-pass procedure, and to some extent it is never totally automatic. Yet, it is precisely on its ability to be indefinitely repeated (and thus verified, and possibly refined) that one of its key strengths rests.

References

AKSIT M., TEKINERDOGAN B. & BERGMANS L. (2001), "The Six concerns for Separation of Concerns", *ECOOP 2001, Workshop on Advanced Separation of Concerns, June 18-22, 2001*, Budapest: Hungary. [http://purl.utwente.nl/publications/37227]

ALLEN W. (1973), *Accent and Rhythm.* Cambridge: Cambridge University Press.

ALLEN W. (1987), *Vox Graeca*, Cambridge: Cambridge University Press.

BLOOMFIELD L. (1914), "Sentence and Word", *TAPhA* 45, 65-75.

BULLOCH A. (1970), "A Callimachean Refinement to the Greek Hexameter", *CQ n.s.* 20, 258-268.

CANTILENA M. (1995), « Il ponte di Nicanore », *in* M. Fantuzzi & R. Pretagostini (eds), *Struttura e storia dell'esametro greco*, Roma : Gruppo editoriale internazionale, 9-67.

DEVINE A. M. & STEPHENS L. D. (1978), *The Greek appositives: toward a linguistically adequate definition of caesura and bridge*, Chicago: The University of Chicago Press.

DEVINE A. & STEPHENS L. (1994), *The prosody of Greek speech*, New York/Oxford: Oxford University Press.

FRÄNKEL H. (1955), *Wege und Formen frügriechischen Denkens*, München : C. H. Beck.

FRUYT M. (1992), « Le mot : aperçu théorique et terminologique », *Lalies* 10, 113-124.

FUSI D. (2002), « Appunti sulla prosodia del Lussorio di Shackleton-Bailey : alcune questioni di metodo », *in* F. Bertini (ed.), *Luxoriana*, Genova : Dipartimento di Archeologia, Filologia classica e Loro Tradizioni, 193-313.

FUSI D. (2004), « Fra metrica e linguistica : per la contestualizzazione di alcune leggi esametriche », *in* E. Di Lorenzo (ed.), *L'esametro greco e latino : analisi, problemi e prospettive – Atti del convegno di Fisciano 28-29 maggio 2002*, Napoli : **editeur**, 33-63.

FUSI D. (2009), "An Expert System for the Classical Languages: Metrical Analysis Components", *Lexis* 27, 25-45.

FUSI D. (2011), *Informatica per le scienze umane*, Roma : Edizioni Nuova Cultura.

GOLDSMITH J. & LARSON G. (1990), "Local Modelling and Syllabification", *in* K. Deaton, M. Noske & M. Ziolkowski (eds), *Proceedings from the 26th Regional Meeting of the Chicago Linguistic Society*, Chicago: Chicago Linguistic Society, 129-141.

LEHRS K. (1865), De Aristarchi studiis Homericis, Lipsiae: S. Hirzekium.

LYONS J. (1968), *Introduction to Theoretical Linguistics*, Cambridge: Cambridge University Press.

MAAS P. (1962), *Griechische Metrik / Greek Metre*, Translated by H. Lloyd-Jones, Oxford: Clarendon Press.

MAYER T. (2010), "Toward a Totally Unsupervised, Language-Independent Method for the Syllabification of Written Texts", *Proceedings of the 11th Meeting of the ACL-SIGMORPHON – ACL 2010*, Uppsala (Sweden): Association for Computational Linguistics, 63-71.

MONTEIL P. (1963), *La phrase relative en grec ancien. Sa formation, son développement, sa structure, des origines à la fin du V^{e} siècle A.C.*, Paris : Klincksieck.

NORBERG D. (1958), *Introduction à l'Étude de la Versification Latine Médiévale*, Stockholm : Almqvist & Wiksell.

O'NEILL (1942)

REICHLER-BÉGUELIN M.-J. (1989), « Conclusion : catégorisation linguistique intuitive et prototypie », *Lalies* 10, 205-214.

ROSSI L. (1963), « Anceps : vocale, sillaba, elemento », *RFIC* 91, 52-71.

SAUSSURE F. de ([1916] 1972), *Cours de linguistique générale*, éd. par C. Bally et A. Sechehaye, Paris : Payot.

SUKHOTIN B. V. (1973), "Deciphering methods as a means of linguistic research", *COLING '73 – Proceedings of the 5th conference on Computational linguistics*, Stroudsburg (PA): Association for Computational Linguistics, 209-214.

THREATTE L. (1980), *The Grammar of Attic Inscriptions*, Berlin/New York: de Gruyter.

VENDRYES (1945)

WEST M. (1982), *Greek Metre*, Oxford: Clarendon Press.

WIFSTRAND A. (1933), *Von Kallimachos zu Nonnos.* Lund: Gleerup.

ABSTRACTS

Daniele Fusi, *Titre en français*

Chiron is a system for the analysis of virtually any language, poetical tradition, metre and text; its modular architecture provides a 2-step analysis process, where data collection is separated from interpretation. The first step happens by chaining any number of analysis layers, each performing a specialized task. In Greek and Latin metrics, a syntax layer is required between the phonological and metrical one for detecting appositives and clitics. Analysis data are saved into a repository for data interpretation; in turn, the observations are stored into a metrics corpus, which can be queried with any complex expression. This provides a sort of live metrics laboratory, especially useful for studying interaction among the complex phenomena underlying metrics.

Keywords: metrics, appositives, expert systems, digital humanities, history of language, metrical corpora, Greek, Latin, Italian

RÉSUMÉS

Daniele Fusi, *A Multilanguage, Modular Framework for Metrical Analysis: It Patterns and Theorical Issues*

Chiron est un système d'analyse de la tradition poétique, du mètre et du texte, s'appliquant virtuellement à n'importe quelle langue ; son architecture modulaire fournit un processus d'analyse en deux étapes, au cours desquelles la collecte des données est séparée de l'interprétation. La première étape consiste à enchaîner un nombre quelconque de couches d'analyse, exécutant chacune une tâche spécialisée. Dans la métrique grecque et latine, une couche syntaxique est nécessaire entre la couche phonologique et la couche métrique pour détecter les appositives et les clitiques. Les données analysées sont enregistrées dans un dépôt en vue de leur interprétation ; à leur tour, les observations sont stockées dans un corpus de métrique, qui peut être interrogé avec une expression complexe. Cela fournit une sorte de laboratoire des métriques attestées, particulièrement utile pour l'étude de l'interaction entre les phénomènes complexes sous-jacents à la métrique.

Mots-clés : métrique, appositive, systèmes experts, humanités numériques, histoire de la langue, corpus métriques, Grec, Latin, Italien