Schriften des Instituts für Dokumentologie und Editorik — Band 13

Versioning Cultural Objects
Digital Approaches

edited by

Roman Bleier, Sean M. Winslow

2019

BoD, Norderstedt

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de/ abrufbar.

Digitale Parallelfassung der gedruckten Publikation zur Archivierung im Kölner Universitäts-Publikations-Server (KUPS). Stand 18. Dezember 2019.

2019

Herstellung und Verlag: Books on Demand GmbH, Norderstedt
ISBN: 978-3-7504-2702-0
Einbandgestaltung: Julia Sorouri and Stefan Dumont; Coverbild gestaltet von Vinayak Das Gupta.
Satz: Roman Bleier, Sean M. Winslow und LuaTEX

Towards a Model of (Variant) Readings

Elisa Nury

Abstract

In scholarly editing, more particularly in the context of collating various versions of a text, the definition of a variant reading is crucial. Yet, despite its importance, the meaning of a variant reading is often reduced to a "difference." The reason for such a vague definition is that what makes a variant can largely depend on the field of study: scholars of the Homeric oral tradition will consider different variants from scholars of medieval traditions or early printed texts, or from genetic critics. This contribution will focus on the modelling of a *reading*, arguing that formalizing this concept is necessary in order to define, and thus model, a *variant*. This article will also address digital representation of a reading by focusing on one implementation: the JSON data format used in conjunction with collation programs such as CollateX.

What is a version? In textual criticism, the term *version* may specifically describe a major rewriting of a work, possibly by the author. Here, however, we will consider versions in a broader sense. The critical comparison—or collation—of different versions of one text is a necessary step during the preparation of a text-critical scholarly edition. Each version of the text is recorded in a document—or witness—and consists of readings, i.e., the particular word or words found at a given point in the text. In this context, a version is determined, amongst other characteristics, by the *differences* in the words found in the text, or *variant readings*. Variant readings are important since they provide valuable information regarding how versions are related to each other and how the text evolved through transmission. This article will focus on the modelling of *readings*, arguing that formalizing this concept is necessary to define, and model, variant readings. We will show how reading was a technical term that was used quite consistently through the ages, until it was defined with precision. Then we will establish the basis for a model by selecting important features of textual readings according to the previously examined definitions. These features, such as the textual content (or absence thereof), its size, and location in the text, will be discussed, raising various issues. This article will also address digital representation of a reading by focusing on one implementation: the JSON data format used in conjunction with collation programs such as CollateX. As we will see, the concept of variant readings may depend on the tradition of the text in consideration, and a variant in Homeric epic is different from a variant in a medieval tradition. The concept of variant is also dependent on the purpose of the comparison: a scholar attempting to reconstruct a

Versioning Cultural Objects: Digital Approaches, edited by Roman Bleier and Sean M. Winslow. Schriften des Instituts für Dokumentologie und Editorik 13. Books on Demand, 2019, 3–23.

stemma, or a linguist, may need to examine different variants. Therefore, a model of a reading should make it possible to distinguish different sets of variants depending on the context, and we will examine how the JSON implementation makes it possible with a few examples.

Let us consider the example of figure 1, where four versions of a sentence are aligned. When comparing the sentences of A, B C,and D, some readings can be considered equivalent in all four sentences, such as *The* or *upon*; other readings are different and change the meaning of the sentence: the absence of the adjective *bright* in sentence B, the triplet *star/sun/stars*, and the verbs with different tense (*shines* and *shone*). Finally, some readings are different, but may not alter the sense of the sentence (such as *worlde* and *world* or *sun* and *sunne*). Readings are thus divided between equivalent readings and different readings, and among the different readings a set of readings may be considered significant variant readings (see figure 2).

A	The	bright	star	shines	upon	the	world.
B	The		sun	shines	upon	the	world.
C	The	bright	stars	shone	upon	the	world.
D	The	bright	sunne	shines	upon	the	worlde.

Figure 1: Readings.

In the short collation extract of figure 1, there are four places where differences appear in the text. However, not all differences between readings are necessarily considered variant readings in any possible context. Scholarly opinions on this point range widely: from the view that every difference is a variant (Andrews) to considering only a limited number of "significant" differences to be variants, for instance, in the context of New Testament criticism, and therefore it is not enough to define a variant simply as a difference:

> The common or surface assumption is that any textual reading that differs in any way from another reading in the same unit of text is a "textual variant", but this simplistic definition will not suffice. Actually, in NT textual criticism the term "textual variant" really means—and must mean—"*significant*" or "*meaningful* textual variant" (Epp 48).

In fact, the concept of *variance* has evolved with time and according to several theories. Since the nineteenth century, many scholars contributed to the development of a

method for the establishment of genealogical relationships between manuscripts: the so-called Lachmann method. Maas in particular focused on a specific category of differences: shared errors, or indicative errors, can be used as a guide in order to assess the witnesses of the text and determine their relationships into a *stemma codicum*, or genealogical tree of textual witnesses.[1]

Greg separated variant readings into accidental and substantial, following the idea that some differences (substantials) have more importance than others (accidentals):

> [W]e need to draw a distinction between the significant, or as I shall call them, *substantive* readings of the text, those namely that affect the author's meaning or the essence of his expression, and others, such in general as spelling, punctuation, word-division, and the like, affecting mainly its formal presentation, which may be regarded as the accidents, or as I shall call them, *accidentals* of the text (Greg 21).

In the twenty-first century, scholars started to compare textual variants to DNA mutations and applied concepts from evolutionary biology and phylogenetics to textual criticism (Barbrook et al.; Salemans; Heikkilä). Lastly, in opposition to the distinction between accidental and substantial variants, Andrews suggested a big data approach where every difference is a variant.

With the introduction of Lachmann's method, shared errors became the object of scholarly attention, and much work was done on the description and classification of the kind of errors committed by scribes who were copying manuscripts by hand. The cause of the error, as well as its conscious or unconscious character, is generally taken into account. Since the conscious modifications of scribal corrections were often attempts at improving or restoring the text, the terms *innovation* and *secondary reading* are frequently preferred to *error*. One of the most comprehensive review of errors was published by Havet, but other scholars have proposed other typologies of errors (Petti; Love; Reynolds and Wilson). These typologies often divide errors into four types: additions, omissions, substitutions and transpositions (Petti). When the scribe is consciously modifying the text, Petti (28–29) refers to scribal corrections as insertions, deletions and alterations instead of additions, omissions and substitutions. In parallel, many fields of study have offered their own definitions for variants according to their needs and their perspective on the text. From oral traditions such as Homeric epic to early printing, from medieval traditions to genetic criticism, from linguistics to phylogenetics, variants take many forms depending on the context: *multiformity* (Nagy), *early* or *late* states (Dane), variants at the sentence level (Cerquiglini), *open* variants, *type-2* variants (Salemans), and so on. The task of proposing a model for

[1] Witnesses are documents which bear a copy of a text, and may be either manuscripts or printed editions. The stemma is a diagram that represents the relationships between those witnesses.

variant readings which would be suitable in any of the possible contexts, seems at best challenging, if not impossible. Rather than dealing directly with variants, this article will focus on modelling readings, especially textual readings. Not all readings are variant readings, but variants are always readings which differ in some respect from one another (see figure 2). Once readings have been modelled, variant readings could be more easily modelled as a set of readings, with various criteria according to each discipline (V1, V2, V3). However, modelling those subsets will not be in the scope of this article. In order to propose a model for readings, we will first review the origins and usage of the term as well as its definitions in Section 1. The analysis of definitions will provide a first outline for a model, which will be discussed in Section 2.

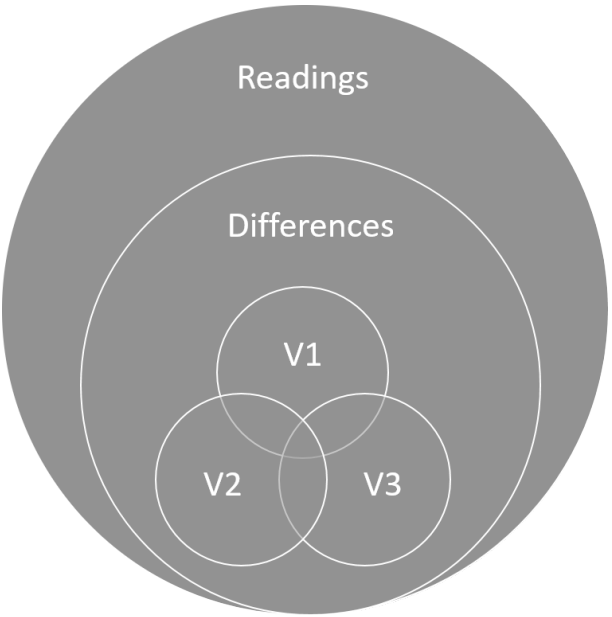

Figure 2: Readings, differences and variants.

1 Readings in context

Reading is a technical term that has long been used in the context of textual criticism and philology. It was already attested with Alexandrian critics: terminology included *graphe* (what is written), and *anagnosis* (what is read, a reading). The Latin equivalents are *scriptura* and the most common *lectio* (Montanari 26). The terms used by scholars

of Antiquity imply a distinction between the words that are actually written on the page as opposed to the interpretation of the text. In English as well, a reading implies a form of interpretation; it could be read in more than one way. Here are a couple of examples where the words *scriptura* and *lectio* are used to qualify textual variation:

> *Obolus, id est, virgula iacens, adponitur in verbis vel sententiis superflue iteratis, sive in his locis, ubi* ***lectio*** *aliqua* ***falsitate*** *notata est, ut quasi sagitta iugulet supervacua atque falsa confodiat.* Isidore 1.21.3.

> The obelus, that is, a horizontal stroke, is placed next to words or sentences repeated unnecessarily, or by places where some **passage** is marked as **false**, so that like an arrow it slays the superfluous and pierces the false. (Barney et al.)

> *"Et idcirco inportunissime," inquit, "fecerunt, qui in plerisque Sallusti exemplaribus* ***scripturam*** *istam* ***sincerissimam*** *corruperunt."* Aulus Gellius 20.6.14.

> "And therefore," said he, "those have acted most arbitrarily who in many copies of Sallust have corrupted a **thoroughly sound reading**." (Rolfe)

Here the nouns *scriptura* and *lectio* have been emphasized, as well as the term which qualifies them. As these passages demonstrate, there was a strong focus in Antiquity on whether a reading is corrupt or sound. When producing a new literary book, Hellenistic scholars used to correct a single copy of a work, instead of comparing as many copies as possible as modern editors do. This practice led Hellenistic scholars to become correctors of a specific work, and some experts compared them to editors (Montanari). Therefore, the need to distinguish between authentic and spurious readings arose, which may have motivated the dichotomy between sound versus corrupt readings, true versus false. The concept of variant reading, however, appeared much later during the Renaissance. In the Renaissance, Humanist scholars who were rediscovering and editing classical texts of Latin and Greek literature started to deploy technical terms that would become the base of the language of textual criticism. Silvia Rizzo's *Lessico Filologico degli Umanisti* provides invaluable information about the vocabulary in use amongst famous Humanists in the fourteenth and fifteenth centuries. By analysing their correspondence and publications, Rizzo was able to extract global definitions and explain what they meant when they used a given word. During the Renaissance, as Rizzo (209–13) shows, *lectio* and *scriptura* continued to be used as synonyms in much the same way as in Antiquity, for a passage of a text that can be read in a manuscript or an edition. Renaissance scholars would apply the term to readings from manuscripts as well as conjectures by other Humanists, and would

mostly describe those readings as either correct (*recta, sincera*) or incorrect (*corrupta, mendosa*) according to their judgement. At the same time, the concept of *variant reading* started to be used more precisely with *varietas* (diversity) and in expressions where *lectio* or *scriptura* were used in connection with the adjective *varius*. Lorenzo Valla and Girolamo Avanzi have both used *varia lectio* and *varia scriptura* to describe a portion of text with different possible readings, as reported by Rizzo (213). Valla was accused by Poggio of having presumptuously corrected a verse from Sallustius' first Elegy. Valla replied to Poggio that he did not emend Sallustius but merely chose one reading in a passage that varies (*varia scriptura*), even though the reading was attested only in very few manuscripts.[2] Another scholar, Avanzi, was asked for his opinion on a difficult passage from Catullus I, 9. He offers no solution of his own to emend the corrupted text, but he sends to his correspondent a list of conjectures (*varia lectio*) proposed by others.[3]

The usage of *lectio* and *scriptura* illustrates two contrasting approaches to readings and variant readings. Usually, a reading becomes a variant only when compared to another reading (Froger 80); variant also implies a deviation from a norm, one version of the text which may be chosen at random (Colwell and Tune 253).[4] On the other hand, a variant can be one among multiple possible alternatives, in a place where at least two witnesses disagree as to what the text is. Consequently, Colwell and Tune decided to refer not to variants, but to *variation-units*. This approach is shared by genetic criticism, which reject the existence of an invariant text, against which variant readings are compared (Biasi). In the twentieth century, formal definitions of reading can be found for instance in editing manuals, dictionaries or lexicons. Stussi defines a reading as "a passage from a transmitted text as it appears in a given witness"(Stussi 89).[5] A more precise definition of a reading is given by Froger, while describing one of the first examples of collation software:

> The form or content of the text in a given place is a *reading*, that is to say what we read at this location. Any manuscript, for instance the original, can

---

[2] "*Nam quomodo videri possum emendare Sallustium, qui, incertum est, an sic scriptum reliquerit, ut me tu ais emendare voluisse? Ego tantum* ***ex varia scriptura****, quid mihi satis videatur, pronuncio. At cur praeponis, inquies, illam scripturam, quae in paucioribus codicibus est? Praepono, non ut Sallustius emendem, sed ut admoneam sequendum, quod plurimorum confirmat authoritas.*" (Valla 263). The discussion can be found in Valla's *Antidoti in Pogium*, book I, in the section on Sallustius.

[3] *"non meam, sed variam lectionem accipies illius versus in primo carmine Catulli"* (Avanzi a5v).

[4] Colwell and Tune explain that the "norm" against which variant readings are compared may be different depending on editors: "So what is commonly done in practice? Some particular text is chosen—often at random—for the norm. Either we use a printed text such as the Textus Receptus, sometimes an edition by Tischendorf, Westcott-Hort, or Nestle; or, we may use the text of a particular MS whose textual affinities are already known, e.g., Vaticanus or Alexandrinus" (Colwell and Tune 253).

[5] "Con lezione di un determinato testimone si designa un passo del testo tramandato così come compare in tale testimone" (Stussi 89).

> be considered regarding its content as a collection or set of readings, which are the text elements at various levels: chapter, paragraph, sentence, word, syllable, letter, and even punctuation or accents (Froger 9).[6]

This definition adds more precision: a reading is a textual element ('what is read'), and it can be of various scope, from the smallest punctuation marks to whole chapters. How can these definitions of a reading lead to a first example of a reading model?

2 Modelling a reading

The purpose of data modelling in the Humanities is to describe and structure information about real-world or digital objects in a formal way, so that this information becomes computable (Flanders and Jannidis 229–30) and so that it can be manipulated and queried with the help of a computer in order to answer questions. Ultimately, the purpose of modelling readings is to help determine if two given readings may be considered variant readings in a specific context. Flanders and Jannidis (234) suggest modelling textual variants in a scholarly edition by classifying variants according to some scheme, such as accidental versus substantial, or orthographical versus lexical, which corresponds to a consensus within the community.

As we have seen, however, variants can represent something very different depending on the perspective (stemmatics, linguistics, etc.) and textual traditions (oral, medieval, early printing, and so on); therefore, readings need to be modelled independently of their function in textual criticism, but with enough information to decide what is a variant in those contexts. It may be helpful to consider the distinction between readings and variants in the framework of Sahle's wheel of text model (Sahle 45–49). Readings can be considered as a part of the text as Document (TextD), whereas variants are part of the text as Version (TextF). The text as Version is further divided into subcategories, such as TextK, a canonical representation of the text which aims at identifying the best (true) text. With this framework in mind, the characterization of readings as authentic or corrupt does not make a good model for readings, since it represents rather variants than readings. Therefore, the more recent definitions of readings may provide a better starting point to the model than the true/false distinction previously applied to readings. Models are simplified representations of an object of study, a selection of features among all available (Pierazzo 44–45). From the overview of the term reading provided in the previous section, in particular the

---

[6] "La forme ou teneur du texte en un lieu donné est une «leçon», c'est-à-dire ce qu'on lit à cet endroit. Un manuscrit quelconque, par exemple l'original, peut donc être considéré, quant à sa teneur, comme une collection ou un ensemble de leçons, qui sont les éléments du texte à différentes échelles: celle du chapitre, du paragraphe, de la phrase, du mot, de la syllabe, de la lettre, et même du signe de ponctuation ou des accents" (Froger 9).

definition of Froger and Stussi, features which apply to a reading can be inferred, namely that a reading:

- conveys textual content;
- has a precise location in the text (also referred to as *locus*);
- can occur at any level of the text, and thus have various sizes;
- is transmitted by a witness.

2.1 Issues

These features need to be discussed in more detail. For instance, is it too restrictive to limit a reading to textual content? What about decorations, mathematical diagrams and other non-textual elements? Historians of Greek, Arabic or Egyptian mathematics have acknowledged the need to collate and critically edit mathematical diagrams instead of simply providing corrected figures to fit modern standards. Raynaud created a stemma for the *Epistle on the Shape of the Eclipse* by Ibn al-Haytham, a mathematical treatise from the eleventh century, using the mathematical diagrams present in the text. In order to collate diagrams and apply Lachmann's method of shared errors, Raynaud had to select "characters" from the diagrams, which could be regarded as an equivalent for readings. This suggests that it is possible to define and model readings for mathematical diagrams. It would be different from textual readings, but as important for the comparison of versions from traditions of mathematical texts. Other types of content could include—and are not limited to—visual content such as decorations, illuminations, or artist's sketches (see the contribution of Martina Scholger in this volume, on comparing the sketches of the Austrian artist Hartmut Skerbisch). Musical compositions need as well to be collated and critically edited, however musical readings and variants are quite different from textual readings and variants: for instance, pitch and metrical values are significant features of a musical note to compare (Broude).

Let us focus here on readings as textual content. Other issues arise with gaps and lacunae: can the absence of text, such as an omission, a so-called *lacuna*, be considered a reading as well? It would seem that the absence of text is by definition not a reading. It cannot be read in the witness, even if it can often be defined by the other features listed above (the size of the missing text may be difficult to evaluate in some cases). However, a missing reading may be significant for the manuscript tradition: since a missing passage is difficult to restore by conjecture, a lacuna shared by several witnesses can often be used as a significant error that indicate a relationship between those witnesses (Reynolds and Wilson 213). A lacuna that helps in grouping manuscripts and building the stemma therefore needs to appear in the collation. How should the absence of text be modelled? As a special kind of reading, or separately? In this model, lacunae were included as readings without any content.

Conjectures—reconstructed readings proposed by scholars which are not present in any witness—seem to qualify as readings according to the features listed above. However, one may ask if conjectures are indeed transmitted by a witness. Conjectures are obviously constituted of textual content of a certain size, meant to be read at a certain location; can they nevertheless be considered to be transmitted by a witness when they are published in a scholarly article instead of an edition? According to Greetham, a conjecture "is involved only when an editor reconstructs or creates a reading which is not extant in any of the witnesses" (352). A conjecture is thus a new reading, with no prior witness evidence, but with an established origin that can be traced to a particular scholar or scribe. In this sense conjectures are considered as part of the reading model.

The location of a reading in the text is not as easy to formulate as it seems. It would not be enough, for instance, to number each word, since the count would then be different for every witness. Even a reference system such as the canonical citations for classical texts can have limitations, when it comes to precision at the word level. Citations such as Pliny nat. 11.4.11 or Vergil ecl. 10.69 refer respectively to the *Natural History* of Pliny the Elder, Book 11, Chapter 4, paragraph 11, or Vergil's *Eclogues* 10, verse 69. The minimal text unit here is the paragraph or the verse, not the word, and at some point in the text, there will be chapters or verses with different word numbers. The location in the text can only be accurately expressed after collation has happened and readings have been aligned with each other. Canonical citations have been formalized in digital formats such as DET (Robinson) or the Canonical Text Services (CTS) Data Model (Crane et al.).

Text can be seen as both a conceptual (immaterial) sequence of words and punctuation from which a reader derives meaning and as a material sequence of marks on a document. Readings are also made of marks recorded on a physical document, besides being part of the immaterial text, thus a reading has both a location in the text and a location in the document where it appears. The document location may be rendered with varying degrees of precision: for instance with folio or page number of the witness in which it appears, with an additional line number, or with a very precise set of coordinates for a two dimensional surface on the page.[7] Finally, it is worth asking if different levels of reading (letters, words, sentences and so on) call for different models and how those levels relate to other existing models. For example, how would the letter level relate to the model used by the DigiPal framework Stokes uses to describe letters from a palaeographical point of view? How would the sentence level relate to the treebank model (Haug) used to annotate textual corpora? How would the different levels be linked together, if the intent of the scholar is to collate at different

[7] See, for instance, the TEI P5 Guidelines chapter 11 for representation of primary sources, in particular section 11.1 on digital facsimiles www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html#PHFAX. Accessed 30 Sept. 2017.

levels? Monella, for instance, decided to collate a text from the Latin Anthology at three different levels, which are called graphical (letters, punctuation), alphabetic (the abstract representation of a letter in a particular alphabet) and linguistic (word) levels.

The different levels may certainly be characterized by additional features of their own. Readings at the word level have a specific spelling, and may be abbreviated. Readings at the word level may also have morphological features, such as lemma or part-of-speech properties or even a phonetic transcription. These linguistic annotations could be useful when comparing readings during collation. For instance, words that do not share gender, number, case and lemma could be considered variants. In the case of oral sources, a different pronunciation may be considered a variant. Layout could also be significant in some contexts: the same word written in bold, or italics or in colour could signal a variation. For instance, Caton argues that a transcription loses information when a word originally written in italics, to denote emphasis, is transcribed into Roman font. At the line level in poetry, metrical patterns would be an important feature. At the sentence level, syntactic information about the subject, object, verb and other elements of the sentence may be an important feature. This information could be particularly interesting for the comparison of versions translated in a different language from the original. In principle, the comparison happens always with readings at the same level: letters are not compared to words, or words to paragraphs. It is worth noting, however, that even if the word level is used during collation, it may be that in the result, words will be grouped together to form a new reading at a different level than the word level (a variation unit that falls between the word and sentence levels). Considering the sentences from the fictive witnesses in figure 1, the groups of words *star shines*, *sun shines*, and *stars shone* may be considered as one reading only, for the purpose of studying the collation results. When there are many variations close to each other, it may be difficult to decide how to group words into readings, if they should be grouped at all, and the readings may be different according to different editors. One could decide to group words instead as *bright star*, *sun* and *bright stars*, with the verb as a separate reading.

2.2 Model

In summary, the model could be expressed as in figure 3: readings can either have content or not. In both cases, a reading has the general features outlined above, such as the witness in which it is found, a position both in the text of the witness and the document of the witness, or a level of precision such as the word level. When the content is present, it can be textual content or another type of content such as diagrams or illustrations. The textual content has a second layer offeatures: syntax, morphology, phonetic, layout, and so on. Depending on the level of the textual content, features may differ. At the sentence level, it is possible to describe the relationships

between words or group of words: *The bright star* is the subject of the verb phrase *shines upon you*, a relationship which is more difficult to represent at a word level. The other types of content would have their own features, such as the characters in diagrams described by Raynaud.

On the other hand, readings without content cannot be described with those additional features. There are other concerns regarding an absence of content, or lacunae. First, there are different reasons behind the presence of a lacuna. The missing text could have been present in the manuscript but is no longer readable by scholars, due to damage or missing pages. In other cases, the copyist marked a lacuna explicitly, with a series of dots for instance, because the text was already missing in the witness serving as the exemplar. The scribe may also have left a blank space to be filled later, and which was never completed. In medieval manuscripts, this would happen easily for materials such as titles, initials or coloured text, which were added later often by a different person than the copyist of the main text. In addition, Dillen has demonstrated the importance of distinguishing between several types of lacunae in Beckett's draft manuscripts, such as authorial lacunae as opposed to editorial ones. Lastly, the lacuna may not be perceptible, unless the witnesses are collated. The collation result could then expose in a witness the absence of a reading which was present in at least one other witness. This kind of lacuna does not belong to the reading model, but only to the variant model: a variant arises either if two readings are considered different, or if a reading is compared against an absence of a reading. In figure 1, the absence of *bright* in witness B would have gone unnoticed unless exposed by the collation against the readings in sentences A and C. The reading may be absent because the scribe did not copy it, whether voluntarily or not, or because it was absent altogether from the exemplar. It is then important to distinguish between the reasons behind a lacuna: is the text present but no longer accessible? Is there a mark indicating that the text was already illegible to the copyist? Or is there no evidence? Even if the text is absent from every witness, the presence of a lacuna can be indicated by inconsistencies in the meaning, for metrical or grammatical reasons, or by incomplete content (such as a missing plural "s").

Given two or more readings at a place of variation, the comparison of the reading's features could help to identify in what aspect the readings differ. This comparison could then lead to a decision regarding which perspective those readings become variant readings of. Let us consider pairs of readings from the sentences in figure 1: comparing the features of *stars* and *star* would show a difference in number, plural and singular, but the lemma would indicate that they represent the same word. It would thus be a grammatical difference. The readings *sun* and *star* have a different lemma, and therefore represent a lexical difference. Two words which share all features (lemma, part of speech and so on) and show no other difference than their original written form would represent an orthographical difference, or graphical difference for

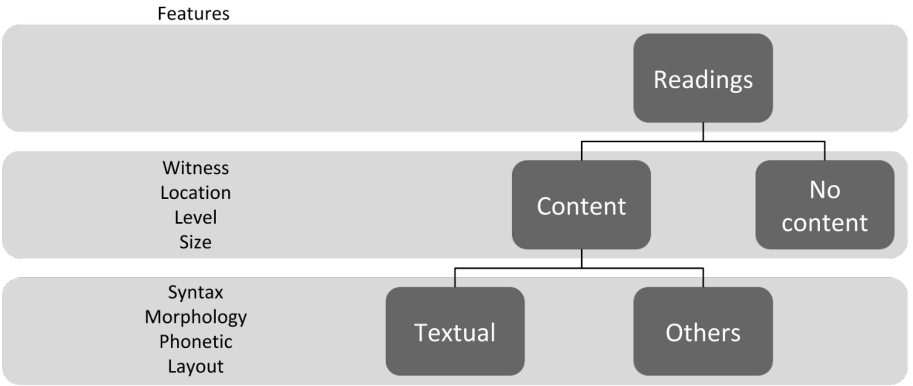

Figure 3: Model for readings.

languages which have no standardized orthography. In different scholarly contexts, the features of readings could be used todefine criteria which are then applied to isolate the relevant variant readings.[8] First, if all differences are considered variants, then readings which display any difference among their features will be considered variants. On the other hand, since orthographical differences are often not considered variants while editing a text (Reynolds and Wilson; Love), the distinction between non-orthographical or orthographical differences allows the editor to select the set of readings which represent grammatical or lexical differences and ignore spelling variants. Finally, linguists would be able to select only spelling variants, particularly significant for the study of language evolution (Vierros and Henriksson). These three contexts will be further examined in Section 4 below, using a practical example. The next section will first deal with the representation of a reading in digital format.

3 Digital representation: from reading to token

To translate the concept of a reading, as defined by centuries of textual scholarship, into digital representation, it seems there is already a counterpart in computational linguistic terminology: the token. Tokens are commonly used for lexical analysis in computer science, as a sequence of characters with an identified *meaning* is converted into a token (see, for instance, Grefenstette and Tapanainen). If manual collation is the comparison of readings, computer-supported collation is the comparison of tokens. Computer-supported collation is the application of computing methods to the

[8] These criteria would not necessarily be applied at the time of recording variants, but also after variants are recorded, to identify only the variants relevant to a specific context.

comparison of textual witnesses: instead of comparing manually the existing versions of a text, digital transcriptions are collated with the help of an alignment algorithm. Juxta and CollateX are two of the most well-known collation tools available, and were both conceived according to the Gothenburg model of collation.

The Gothenburg model was devised in 2009 in order to define computer-supported collation. It divides the process of collation into four successive tasks (Dekkeret al.). The first of these tasks is to split the entire text of each witness into smaller units, called tokens, to be compared. The other tasks include alignment of those tokens (the actual collation process), analysis, and output of the collation results. The parallel between Froger's reading definition (see above) and a token is clear. In the Gothenburg model, a text is divided into a list of tokens which are textual units (a sequence of characters) at a chosen level. This is also how Froger describes a text, as a collection of readings, which are made of the text's content taken at a particular level. As such, the tokens share the same features as readings: the textual content of a witness, with a precise location in the text determined by its position in the full list of tokens, and at a specific level.

According to Dekker et al. (4), a token is a textual unit at "any level of granularity, for instance, on the level of syllables, words, lines, phrases, verses, paragraphs, text nodes in a normalized XML DOM instance, or any other unit suitable to the texts at hand." The CollateX documentation more explicitly considers a token as a textual unit that ideally carries meaning, thus above the character level.[9] At letter level, phenomena such as transposition are much more frequent and reduce the efficiency of the alignment algorithm. For this reason, collation is preferably performed at a higher level, rather than at character level. However useful for the collation process, this restriction does not apply in palaeography where letters are the comparison units. Projects such as Monella's also require analysis at character level. From a theoretical and modelling perspective, it is thus necessary not to make assumption about the meaning of a token. The transcription model of Huitfeldt, Marcoux, and Sperberg-McQueen provides a more adapted description for a token, since they do not make a distinction between tokens as characters, as words, or as other levels.[10]

In digital format, the most basic form of a token is a simple string of characters, a linear sequence of one or more symbols representing letters, but with no linguistic in-

[9] See the CollateX documentation: collatex.net/doc/#tokenization. Accessed 27 Oct. 2016.

[10] "A mark is a perceptible feature of a document (normally something visible, e.g. a line in ink). Marks may be identified as tokens in so far as they are instances of types, and collections of marks may be identified as sequences of tokens in so far as they are instances of sequences of types. In other words, a mark is a token if, but only if, it is understood as instantiating a type. The distinction among marks, tokens, and types may be applied at various levels: letters, words, sentences, and texts" (Huitfeldt, Marcoux, and Sperberg-McQueen 297). The transcription model is "agnostic about whether the types (and tokens) it is concerned with are those at the character level or those at the level of words and lexical items" (Huitfeldt, Marcoux, and Sperberg-McQueen 298).

terpretation attached to them. Nevertheless, collation tools usually offer to normalize tokens in order to minimize what is perceived as insignificant variation: typically, normalization permits the removal of upper case, punctuation or other aspects (such as, for instance, hyphenation or line breaks in Juxta, white space characters in CollateX) from the tokens that will be compared, so that these would not be considered differences: *the* and *The* would be treated as the same word for the purpose of aligning the versions together. However, if this normalized form is not explicitly included in the token, it will not be available in the results of the collation. For example, in the case when accidental differences are not significant, the pair of readings *sun*/*sunne* and *world*/*worlde* may be considered as irrelevant differences and thus should be ignored when searching for semantic variants. However, given only the string of characters it is impossible to discriminate between a significant variant such as *shines*/*shone* and the orthographical variants such as *world*/*worlde*. On the other hand, if the reading *worlde* also includes a normalized form *world*, it is then possible to compare the normalized form of *worlde* and decide that it is equivalent to the reading *world*. As a consequence, it could be extremely difficult to distinguish between orthographical or non-orthographical differences without normalized forms, when analysing collation results.

3.1 Token format in CollateX

CollateX makes it possible to distinguish between the original token and a normalized form provided by the user thanks to an input format in JSON, a lightweight data-interchange format.[11] The structure of CollateX's JSON input is described in the CollateX Documentation (2013). Tokens can therefore be represented as JSON objects with various properties, such as:

- t: the textual content in its original form.
- n: a normalized form of the same textual content.

The normalized form is used to align the texts as accurately as possible, while the original content is still available should it be needed by the user when analysing the results; the JSON format of CollateX is thus a very effective way to represent readings involving textual content. However, the absence of content is problematic, since a token must always have at least a property *t* with a positive value. As a result, it is not possible to collate empty tokens, which is a limitation since lacunae are considered readings in this model and need to be represented as tokens as well. So far, I have represented lacunae present in the text due to damage, or explicitly marked by the copyist, as tokens with the textual content *t* as "…", and the normalized form *n* as

[11] See www.json.org. Accessed 10 Mar. 2017.

"lacuna", a combination of content that does not appear elsewhere in the witnesses and therefore cannot be confused with another reading. Lacunae which are revealed by the collation, because a portion of text was omitted by a scribe, are not represented by a token. Instead, CollateX inserts empty tokens in the collation to compensate for the absence of text (Dekker et al.).

As CollateX is used in other projects, their encoding choices may provide further ideas about the representation of readings as tokens. As an example, the Collation Editor, a tool prepared for the collation of the Greek New Testament with CollateX, provides a description of the token's properties online.[12] The Collation Editor provides two layers of normalization and regularization: the original token is normalized in a first step into *t*, with operations such as setting the words in lower case. Then, the token *t* may be regularized again into *n* according to rules defined by the user, which are provided through a *rule_match* feature.

Besides *t* and *n*, any additional properties can be provided to the token object, but will be ignored during collation. Nevertheless, these additional properties would still be available in the results for further processingsuch as visualization. For tokens at the word level, such properties could also include:

Identification. A way to identify and locate the token in the document where it appears, with a reference to page and line numbers for instance. The location may also help to situate the token in the text (with a reference system, such as canonical citations for classical texts mentioned above). A unique identifier could also serve to link the collation result to the transcription, where other properties of the token are encoded and could be retrieved. The Collation Editor includes properties such as *index*, *siglum*, *verse* and *reading* in order to provide identification for each token.

Markup. XML transcriptions of the witnesses are often used within collation software. Since a lot of valuable information is already encoded in the transcriptions, including layout information, several projects have decided to keep the markup in the token properties. It could be exploited during the collation process: for instance, a word marked as bold could be considered as different from the same word in italics. It could also serve to display tokens with more precision. The Beckett Digital Manuscripts project, for instance, displays additions and deletions thanks to this markup property.[13]

Facsimile. A reference to a digital image, for instance in the form of a link, could be helpful to visualize the original reading in context and assess the transcription accuracy (see Nury).

---

[12] The Collation Editor is a tool produced by The Institute for Textual Scholarship and Electronic Editing (ITSEE) at the University of Birmingham. It is an open source tool available on Github: github.com/itsee-birmingham/collation_editor. Accessed 1 Feb. 2017.

[13] See the update from 17 Sept. 2014 here:www.beckettarchive.org/news.jsp. Accessed 31 Oct. 2016.

Linguistic properties. Linguistic properties could be expressed with a standardized format of detailed linguistic annotation, such as part-of-speech, and morphology. Although Crane argues that morpho-syntactic analysis is one major feature of a digital edition, Monella (184) recognizes that the additional workload may be an issue for the encoder. The use of semi-automatic annotation methods still needs to be explored in further research. Smith and Lindeborg propose to use a "dictionary form" to recognize identical lexical readings, and metrical units to compare the rhythm of Iliadic verses. The use of lemma, synonyms and part-of-speech tagging is also planned to be implemented in collation with the tool iAligner (Yousef and Palladino).

Lacunae. If lacunae are represented as tokens, a description of the lacuna's length and reason (such as damage, or missing pages) could be added. In the Collation Editor, lacunae are not represented as tokens, but are included in the properties of the preceding token: *Gap_after*, a boolean variable set to true, records the presence ofalacuna aftera given token. Another property, *Gap_detail*, gives information about the length of the lacuna.

4 Comparing tokens in different contexts

As described above in Section 2.2, tokens can be compared to find variant readings according to a specific perspective. Three possible situations were taken into account: (a) every difference is a variant, (b) only non-orthographic differences are variants, and (c) only spelling differences are variants. Using the properties *t* and *n* of JSON tokens already make it possible to distinguish variants for these three different contexts. Let us consider again the example of a collated sentence in figure 1. The reading *sunne* was normalized to *sun* and the reading *worlde* was normalized to *world*.

In the first situation, all differences are variant readings. Therefore, in each column, the tokens are compared on the basis of their property *t*: in the first column, all tokens have the same property *t*, *The*, and thus there is no variant. In the second column, the absence of *bright* in witness B is a variant, and so on. When each reading has been examined, the following figure 4 highlights every variant.

In the second scenario, orthographic differences are irrelevant. In order to find the relevant variant readings, the tokens must then be compared on their normalized property *n*, so that orthographic differences appearing in property *t* are ignored. In our example, this means that the last column will not show a variant, because witnesses C and D will have the word *world* as a normalized form: when comparing this normalized form to the tokens in witnesses A and B, there will be no difference. The two tokens show a spelling difference (in property *t*) but are in fact considered the same reading because they share the same property *n*. figure 5 shows non-orthographic variants.

A	The	bright	star	shines	upon	the	world.
B	The		sun	shines	upon	the	world.
C	The	bright	stars	shone	upon	the	world.
D	The	bright	sunne	shines	upon	the	worlde.

Figure 4: Every difference is a variant.

A	The	bright	star	shines	upon	the	world.
B	The		sun	shines	upon	the	world.
C	The	bright	stars	shone	upon	the	world.
D	The	bright	sunne	shines	upon	the	worlde.

Figure 5: Non-orthographic differences are variants.

Finally, orthographic variants can be isolated when searching for tokens which share the same normalized form *n*, but not the same original form *t*. In our example, there are thus two columns which contain an orthographic variant (see figure 6). The table could then be reduced to a list of orthographic variants only:

1. sun (B) – sunne (D)
2. world (ABC) – worlde (D)

These three simple examples are of course generalizations: in reality, the principles of collation maybe far more complex. For example, spelling differences may be ignored, except in proper nouns (Love 52). In some cases, it may be difficult to distinguish between a spelling difference or a morphological one. In addition, different readers may give diverse interpretations for certain words or sentences, as it is the case with annotated treebanks (Bamman and Crane). Uncertainty and multiple interpretations thus need to be represented as well. However, if the tokens contain more detailed information, it may help to bring more precision when deciding which readings should be considered as variant readings.

A	The	bright	star	shines	upon	the	world.
B	The		sun	shines	upon	the	world.
C	The	bright	stars	shone	upon	the	world.
D	The	bright	sunne	shines	upon	the	worlde.

Figure 6: Orthographic differences are variants.

5 Conclusion

Different versions of a text are characterized in part by their variant readings. To represent variant readings in digital format, it may be helpful to precisely define and formalize the concept. What is a variant reading, however, is highly dependent on the tradition in question (oral, medieval, early print, etc.) and on the scholarly perspective on the text (stemmatics, linguistics, and so on, following Sahle's wheel of text for instance). As a result, the set of differences present in a textual tradition are not all considered significant in every situation; variant readings are only a subset of all the differences, and different contexts call for different sets of variant readings, as we have seen in the last section.

A first step in formalizing variant readings may be to model and formalize *readings*, in such a way that later, those readings can be compared efficiently in order to define which readings are considered to be variant readings in a given context. The definitions of the term reading thus provided a series of features which can be used to create a model of a reading. However, those features raised a few issues regarding their content, their position in the text as well as in the document, and their relationship between different levels of reading (from characters to words, sentences, and so on). Following the discussion on these issues, a model was proposed that distinguishes between readings with content or without content. The readings with content can again be divided according to the type of content, such as textual or non-textual. The translation of *readings* to *tokens*, using the CollateX JSON format, showed how the use of a simple normalized form could allow to find different sets of variants in practice, within collation results, according to three different contexts. However, as more information is associated with a reading, it could be possible to define variant readings even more precisely. The aim of the model is to represent readings independently of their function in textual criticism, but with enough information so as to decide when a difference becomes a variant. Considering other sorts of content,

such as mathematical diagrams, images or music, the model is flexible enough to for future extension to incorporate other types of content, such as non-textual readings as well.

Bibliography

Andrews, Tara. "The third way: philology and critical edition in the digital age." *Variants*, vol. 10, 2012, pp. 61–76.

Avanzi, Girolamo. *In Val. Catullum et in Priapeia emendationes.* Tacuinus, 1495.

Bamman, David, and Gregory Crane. "The Ancient Greek and Latin Dependency Treebanks." *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, 2011, pp. 79–89.

Barbrook, Adrian C. et al. "The phylogeny of *The Canterbury Tales*." *Nature*, vol. 394, 1998, p. 839. *Astrophysics Data System*, adsabs.harvard.edu/abs/1998Natur.394..839B. Accessed 25 Oct. 2013.

Beckett, Samuel. *Digital Manuscript Project.* 2016. www.beckettarchive.org. Accessed 12 Nov. 2017.

Broude, Ronald. "When Accidentals are Substantive: Applying Methodologies of Textual Criticism to Scholarly Editions of Music." *Text: Transactions of the Society for Textual Scholarship*, vol. 5, 1991, pp. 105–20.

Caton, Paul. "Lost in Transcription: Types, Tokens, and Modality in Document Representation." *Digital Humanities 2009. Conference Abstracts. University of Maryland, College Park June 22–25, 2009.* Maryland Institute for Technology in the Humanities (MITH), 2009, pp. 80–2.

Cerquiglini, Bernard. *Éloge de la variante: Histoire critique de la philologie.* Seuil, 1989.

CollateX. The Interedition Development Group, 2010-2017. collatex.net. Accessed 17 July 2014.

Colwell, Ernest Cadman, and Ernest W. Tune. "Variant readings: classification and use." *Journal of Biblical Literature*, vol. 83, no. 3, 1964, pp. 253–61. *JSTOR*, www.jstor.org/stable/3264283. Accessed 15 June 2015.

Crane, Gregory. "The Digital Loeb Classical Library - a view from Europe." *Perseus Digital Library Updates*, 2014. sites.tufts.edu/perseusupdates/2014/09/22/the-digital-loeb-classical-library-a-view-from-europe. Accessed 31 Oct. 2016.

Crane, Gregory et al. "Cataloging for a billion word library of Greek and Latin." *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage - DATeCH '14*, pp. 83–8, 2014. *ACM Digital Library*, dl.acm.org/citation.cfm?id=2595188.2595190. Accessed 12 Apr. 2016.

Dane, Joseph A. "The Notion of Variant and the Zen of Collation." *The Myth of Print Culture: Essays on Evidence, Textuality and Bibliographical Method*, edited by Joseph A. Dane, University of Toronto Press, 2003, pp. 88–113.

De Biasi, Pierre-Marc. *La génétique des textes.* Natan, 2000.

Dekker,Ronald, et al. "Computer-supported collation of modern manuscripts: CollateX and the Beckett Digital Manuscript Project." *Literary and Linguistic Computing*, 2015, vol. 30, no. 3, pp. 452–70. doi.org/10.1093/llc/fqu007. Accessed 28 Oct. 2016.

Dillen, Wout. "'(Hiatus in Ms.)'. Towards a TEI compliant typology of textual lacunae in Samuel Beckett's manuscripts." *Manuscrítica. Revista de crítica genética*, vol. 28, 2015, pp. 65–73.

Epp, Eldon. J. "Towards the Clarification of the Term 'Textual Variant'." *Studies in the Theory and Method of New Testament Textual Criticism*, edited by Gordon D. Fee and Eldon J. Epp, Eerdmans Publishing, 1993, pp. 47–61.

Flanders, Julia, and Fotis Jannidis. "Data Modeling." *A New Companion to Digital Humanities*, edited by Susan Schreibman et al., Blackwell Reference online, 2015, pp. 229–37. www.blackwellreference.com/public/tocnode?id=g9781118680643_chunk_g978111868064318. Accessed 5 Mar. 2017.

Froger, Jacques. *La critique des textes et son automatisation.* Dunod, 1968.

Greetham, David. *Textual Scholarship: An Introduction.* Garland, 1994.

Grefenstette, Gregory and Pasi Tapanainen. "What is a word, what is a sentence? Problems of tokenization." *COMPLEX 1994: 3rd conference on computational lexicography and text research, Budapest, Hungary, 7-10 July*, 1994, pp. 79–87.

Greg, Walter W. "The Rationale of Copy-text." *Studies in Bibliography*, 1950-51, vol. 3, pp. 19–36.

Haug, Dag. "Treebanks in historical linguistic research." *Perspectives on Historical Syntax*, edited by Carlotta Viti, Benjamins, 2015, pp. 188–202.

Havet, Louis. *Manuel de critique verbale appliquée aux text es latins.* Librairie Hachette, 1911. *Internet Archive*, archive.org/details/manueldecritique00haveuoft. Accessed 5 July 2015.

Heikkilä, Tuomas. "The Possibilities and challenges of computer-assisted stemmatology: the example of Vita et miracula s. Symeonis Treverensis." *The Analysis of Ancient and Medieval Texts and Manuscripts: Digital Methods*, edited by Tara Andrews and Caroline Macé, Brepols, 2014, pp. 19–42.

Huitfeldt, Claus, et al. "What is transcription?" *Literary and Linguistic Computing*, vol. 23, no. 3, 2008, pp. 295–310.

Juxta Commons. The Nineteenth-century Scholarship Online (NINES), University of Virginia. www.juxtacommons.org. Accessed 27 Oct. 2016.

Love, Harold. "The Ranking of Variants in the Analysis of Moderately Contaminated Manuscript Traditions." *Studies in Bibliography*, vol. 37, 1984, pp. 39–57. *JSTOR*, www.jstor.org/stable/40371792. Accessed 21 Oct. 2013.

Maas, Paul. *Textual criticism.* Translated by Barbara Flower. Clarendon Press, 1958.

Monella, Paolo. "Many witnesses, many layers: the digital scholarly edition of the *Iudicium coci et pistoris* (Anth.Lat. 199 Riese)." *Digital Humanities: Progetti Italiani Ed Esperienze Di Convergenza Multidisciplinare. Atti Del Convegno Annuale Dell'Associazione Per L'Informatica Umanistica E La Cultura Digitale (AIUCD) Firenze, 13–14 Dicembre 2012*, edited by Fabio Ciotti, Quadernidigilab, 2014, pp. 173–206.

Montanari, Franco. "From Book to Edition: Philology in Ancient Greece." *World Philology*, edited by Sheldon Pollock et al., Harvard University Press, 2015, pp. 25–44.

Nagy, Gregory. "The Homer Multitext." *Online Humanities Scholarship: The Shape of Things to Come. Proceedings of the Mellon Foundation Online Humanities Conference at the University of Virginia March 26–28, 2010*, edited by Jerome McGann et al., Rice University Press, 2010, pp. 87–112.

Nury, Elisa. "Visualizing Collation Results." *Advances in Digital Scholarly Editing*, edited by

Peter Boot et al., Sidestone Press, 2017, pp. 317–31.

Petti, Anthony E. *English Literary Hands from Chaucer to Dryden.* Edward Arnold, 1977.

Pierazzo, Elena. *Digital Scholarly Editing: Theories, Models and Methods.* Ashgate, 2015.

Raynaud, Dominique. "Building the stemma codicum from geometric diagrams." *Archive for History of Exact Sciences*, vol. 68, no. 2, 2014, pp. 207–39.

Reynolds, Leighton D., and Nigel Guy Wilson. *Scribes and scholars. A guide to the transmission of greek and latin literature.* 3rd ed. Clarendon Press, 1991.

Rizzo, Silvia. *Il lessico filologico degli umanisti.* Edizioni di storia e letteratura, 1973.

Robinson, Peter. "Some principles for making collaborative scholarly editions in digital form." *Digital Humanities Quarterly*, vol. 11, no. 2, 2017. www.digitalhumanities.org/dhq/-vol/11/2/000293/000293.html. Accessed 10 Oct. 2017.

Sahle, Patrick. *Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 3: Textbegriffe und Recodierung.* Books on Demand, 2013. kups.ub.uni-koeln.de/5353. Accessed 3 Nov. 2014.

Salemans, Ben. *Building Stemmas with the Computer in a Cladistic, Neo-Lachmannian, way: the case of fourteen text versions of lanseloet van denemerken.* Katholieke Universiteit Nijmegen, 2000.

Smith, David Neel, and Stephanie Lindeborg. "Comparing Digital Scholarly Editions." *Digital Humanities 2016: Conference Abstracts.* Jagiellonian University & Pedagogical University, 2016, pp. 686–7.

Stokes, Peter. "Modeling Medieval Handwriting: A New Approach to Digital Palaeography." *DH2012 Book of Abstracts*, edited by Jan Christoph Meister et al., University of Hamburg 2012, pp. 382–5. www.dh2012.uni-hamburg.de/conference/programme/abstracts/modeling-medieval-handwriting-a-new-approach-to-digital-palaeography/. Accessed 23 March 2017.

Stussi, Alfredo. *Introduzione agli studi di filologia italiana.* Il Mulino, 1994.

TEI Consortium. "11.1 Digital Facsimiles." *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* Version 3.2.0. Last updated on 10th July 2017. www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html#PHFAX.

The Etymologies of Isidore of Seville, edited and translated by Stephen A. Barney et al., Cambridge University Press, 2006.

Valla, Lorenzo. *Laurentii Vallae Opera.* Apud Henricum Petrum, 1540.

Vierros, Marja, and Erik Henriksson. "Preprocessing Greek Papyri for Linguistic Annotation." forthcoming, *Archive ouverte HAL*, https://hal.archives-ouvertes.fr/hal-01279493. Accessed 8 Feb. 2017.

Yousef, Tariq, and Chiara Palladino. "iAligner: A tool for syntax-based intra-language text alignment." *Fifth AIUCD Annual Conference*, Venice, 2016, pp. 201–5.