Journal of the Text Encoding Initiative

Issue 11 | July 2019 - June 2020
Selected Papers from the 2016 TEI Conference

Encoding of Variant Taxonomies in TEI

Helena Bermúdez Sabel

Electronic version
URL: http://journals.openedition.org/jtei/2676
DOI: 10.4000/jtei.2676
ISSN: 2162-5603

Publisher
TEI Consortium

Electronic reference
Helena Bermúdez Sabel, « Encoding of Variant Taxonomies in TEI », *Journal of the Text Encoding Initiative* [Online], Issue 11 | July 2019 - June 2020, Online since 19 December 2019, connection on 01 July 2020. URL : http://journals.openedition.org/jtei/2676 ; DOI : https://doi.org/10.4000/jtei.2676

Encoding of Variant Taxonomies in TEI

Helena Bermúdez Sabel

ABSTRACT

The inherent flexibility of the digital format has favored the rise of editions that enable access to every witness of a particular textual work. These types of editions might have different goals and seek to answer different research questions, but they usually coincide in drawing attention to the importance of textual variants. To maximize the computational analysis that may be practiced with the variants in different witnesses, a complex taxonomy that reflects the diversity of cases is required.

Many scholars have followed the recommended TEI method for encoding types of variants—that is, through the attributes `@cause` or `@type` inside the element `<rdg>`—while others find that method insufficient. These attributes are not able to enclose the hierarchy intrinsic to complicated taxonomies or the overlap of classes in an efficient way. However, the TEI Guidelines do offer a module that addresses this complex encoding issue: feature structures. The method proposed in

this paper does not advocate for a controlled vocabulary to categorize types of variants. What it offers instead is a pliable encoding method that allows the editor to include multiple layers of information in each apparatus tagset.

INDEX

Keywords: feature structures, textual variants, scholarly editions, taxonomies

ACKNOWLEDGEMENTS

This work was supported by the Ministerio de Educación, Cultura y Deporte (Spanish Government) as part of its predoctoral training program (FPU grant AP2012–4518) and by the Ministerio de Economía, Industría y Competitividad as part of the project *Paleografía, Lingüística y Filología. Laboratorio on-line de la lírica gallego-portuguesa* (FFI2015–68451–P).

1. Introduction

1 This paper proposes a method for introducing multilayered analytical information as part of the critical annotation of an edition. Section 2 gives a brief introduction about the importance of variation studies, justifying the subject of analysis. The method, explained in sections 3 and 4, explores the use of feature structures for the creation of complex taxonomies describing textual variation. Section 5 offers examples of one of the ways in which the variant taxonomy may be linked to the body of the edition.

2 Although this paper is TEI-centered, other XML technologies will be mentioned. Section 4 includes a brief commentary on using XSLT to transform a TEI-conformant definition of constraints into schema rules. However, the greatest attention to an additional technology is in section 6, which discusses the use of XQuery to retrieve particular *loci critici* and to deploy quantitative analyses.

2. Rationale

3 In textual scholarship, two founding concepts shape the methodology of analysis: text and variant. While *text* is a broad concept that requires a theoretical approach to define it with precision, the concept of *variant* is intrinsic to the philological tasks involved in the editorial process, that is, the

empirical operations that define what constitutes a variant and its role in the analysis of the text: normalizations, the selection of contents of the critical apparatus and how this information will be structured, and explanatory annotations, among others.

4 The amount of data we can collect from the analysis of variation depends on the complexity of the transmission process. Scholars who study textual traditions transmitted by scribes are familiar with the use of variants to help identify how many hands participated in the creation of the witness, and where these people came from or where they were trained. Furthermore, variants account for the modifications of the text that may help clarify the political and cultural context in which a work is copied, edited, and published. In addition, variants are also of great importance in modern and contemporary works: the typographer, the editor or editors, and the author at different stages of the composition process must also be studied through the analysis of variants (Segre 1995).

5 The most common issues addressed through the study of variation include the following:

- The analysis of the results of the collation of witnesses is a mandatory step to establish the relationships of filiation among them and, consequently, the construction of the *stemma*—when possible—or local *stemmata* in highly contaminated traditions (Mink 2000).
- Stemmatic analysis sheds light on stages of transmission that are not preserved directly, but that can nonetheless be identified as sources for the extant witnesses—that is, it helps in the description of an archetype and hyparchetypes (Maas 1958).
- The exploration of internal variants may reveal different stages during which the textual materials that form a work were compiled.
- Alongside codicological analysis, variants may contribute information about the time and geographical location in which a particular witness was created.
- Variants present data of interest for the development of the history of writing and the history of a language.
- Variants may provide evidence for the identification of copyists, historical owners, editors, and other persons who may have participated in the transmission of a work or in the alteration of its contents. In addition, variants may provide sociohistorical information through the study of the motivation underlying those alterations.
- The study of variation is a key feature of genetic editions that allows one to analyze the creative process of a particular writer or editor.

6 Considering all of the above, it goes without saying that any type of scholarly edition that takes into account a multi-witness tradition needs to refer to variation somehow. How these references will be made depends on the theoretical background and the editorial model. Textual scholarship is very rich in its treatment of the relation of variants and the final result of the edited text, but both "materialistic" approaches to the edition of texts (for example, Cerquiglini 1989) and methods that look for a curated version (for example, Chiesa 2002) require the analysis of variants.

7 The information a scholar might want to retrieve from the analysis of variants is so dependent on the circumstances of the text and on their interests that a generic taxonomy able to enclose the complexity of variation and reflect all the nuances is impractical. Thus, the goal of this paper is to provide an encoding mechanism that would formalize, for all intents and purposes, textual variants.

8 Among the studies concerning textual variation that can be conducted through a taxonomy of variants, there is the possibility of combining both quantitative and descriptive approaches.

9 A fine categorization of the types of textual variants enriches the critical annotation of any scholarly edition because of the ways in which this information can be embedded in a digital environment. In addition, the analysis of a work after the ordering and categorization of all variants may open a window onto new research questions. For example, a classification regarding content—that is, addition, omission, and mutation—can be systematized according to subdivisions that offer possible explanations for the variation. Applying this kind of analysis to the oeuvre of an author or authors could shed light on the composition process, as well as serve as a source for different types of genetic and stylometric studies.

10 When the object of study is a textual tradition transmitted by acts of copying, a multilayered analysis is especially appropriate. The analysis of linguistic variation between witnesses brings to attention the *core*, original language of the text, and the *patina*, the linguistic layers left by the copyists (Trovato 2014). In addition, linguistic variants shed light on the distinctive features of every witness, which makes the study of linguistic aspects of the textual tradition as a whole more practicable. Furthermore, linguistic variants may occur as a result of self-dictation by copyists after reading and memorizing the extracts they intend to copy, and thus may reflect features of their idiolects. The study of those features contributes to a better understanding of the history of a witness, since it provides evidence of the agents that participated in its creation.

11 Additionally, the linguistic data that can be retrieved by the classification of linguistic variants are of interest to historical linguistics in ways that may transcend the critical analysis of a particular textual work. Any linguistic variant must be analyzed first in a synchronic context: the oldest variants coexist with newer variants in the same linguistic community, but in different contexts or registers (Jakobson 1971, 528). Thus, by identifying linguistic variants in a multi-witness tradition, we are discerning the variants that are competing in a specific chronological framework. Historical corpora used for linguistic studies have sometimes relied on scholarly curated texts or on a transcription of only the oldest witness, and in those cases linguistic variation within the tradition may be neutralized.[1]

12 Depending on the historical period to which a tradition belongs, even graphic variants, often dismissed as nonsignificant variation because they may not be *textually* significant, may nonetheless provide a rich source of cultural information. A graphemic analysis derived from a taxonomy allows an efficient computation of frequencies that can be filtered according to different codicological or historical aspects encoded in the corpus, such as the quire or folio number. This analysis may clarify certain elements of the genesis of the witnesses, such as the identification of hands or editorial interventions. Similarly, scribal errors may provide information related to the sources, such as the distinctive features of a hyparchetype (for example, an unusual graphic substitution might mean that the affected letters had similar shapes in the model) or linguistic data.

3. Feature Structures

13 From the TEI documentation (see the set of attributes specific to elements representing variant readings: TEI Consortium 2016, Appendix B: Attribute Classes, "att.textCritical") we can assume that the recommended way to describe the motivation behind a variant and its categorization is done through the attributes `@type` and `@cause`. However, the complexity of variation, with overlapping categories and complex hierarchies of variation types, cannot be recorded with a straightforward use of these two attributes.[2]

14 The TEI offers a module well suited for the encoding of complex taxonomies: the Feature Structures module (TEI Consortium 2016, 18). A feature structure is a group of `attribute:value` pairs, where the values may be either atomic or nested feature structures. As described by Witt and Stegmann (2009), feature structures are a generic method to organize data with a metarepresentation format that presents numerous advantages, some of which will be discussed in greater detail below. For a more detailed description of feature structures and their rationale see Pose, Lopez, and Rosemary (2014, 9–10).

15 One argument in support of the use of feature structures is the official recognition of the model in 2012 as an international standard (ISO 24612) (Romary 2015), which bestows a certain stability on the methodology and confirms, at least to some extent, its importance and influence within a user community.

16 With respect to the model itself, by describing an object as a sum (and convergence) of multiple individual features, we can implement a complex description with a fine level of granularity. See, for example, in example 1 a proposed definition of a phenomenon of *progressive nasalization*, a type of linguistic variation characterized by the addition of a nasal sound.

Example 1. Entry in the taxonomy for a particular linguistic phenomenon.

```
<fs xml:id="pro-nas">
 <f name="taxonomy">
  <fs type="linguistic">
   <f name="category">
    <fs type="phonetic">
     <f name="process">
      <fs type="alteration">
       <f name="mode">
        <fs type="assimilation">
         <f name="procedure">
          <symbol value="nasalization"/>
         </f>
         <f name="direction">
          <symbol value="progressive"/>
         </f>
        </fs>
       </f>
      </fs>
     </f>
     <f name="position">
      <symbol value="end"/>
     </f>
     <f name="constriction">
      <binary value="true"/>
     </f>
    </fs>
   </f>
  </fs>
 </f>
 <f name="description">
  <string xml:lang="en">Progressive nasalization</string>
```

```
    </f>
  </fs>
```

17 Let us briefly describe the encoding presented in example 1 by going through the different layers of data employed to describe this type of variant. This and other samples presented throughout this paper are part of a dissertation project that studies linguistic variation phenomena between two different manuscript branches that transmit secular Galician-Portuguese medieval poetry. For that reason, the examples used here reflect the research questions of that project, which are related to the study of variation as a source of information for the genesis of the different witnesses.

18 The two features that primarily describe the variant presented in example 1 are the taxonomy that will be used (in this case, the linguistic one) and a feature called `"description"` that contains a definition of the phenomenon.

19 The category of the linguistic taxonomy to which this variant belongs is the phonetic one, and the phenomenon of progressive nasalization is further defined by the features `"process"`, `"position"`, and `"constriction"`. As we can see, the `"process"` feature is more complex than the others and it requires further decomposition. The selection of features to define the phenomenon was based on their relevance for the work in which this taxonomy is applied. For example, in medieval poetry it is important to know which phonetic phenomena occur at the end of a word creating a consonantic coda (constriction): the metrical analysis of the medieval Galician-Portuguese tradition depends on the number of syllables, and progressive nasalization can alter that count through a literary device known as *synalepha*. A synalepha is the merging of two syllables into one whenever a word ends in a vowel and the next word also begins with a vowel. This means that progressive nasalization, with the addition of a nasal consonant, would prevent that merging. If one witness presents a regular metric paradigm and in the other the paradigm is broken because of the presence or the absence of this phenomenon, it is the deviant witness that presents a "spurious" variant.

20 One of the advantages of the internal organization of feature structures is that any layer of information may be referenced during the description of the value of other features. This facilitates the creation of long and complex taxonomies on the foundation of a small set of shared features. For example, in the case of a linguistic variation taxonomy, it could be convenient to define a feature structure that would represent any phonetic phenomenon that implies the addition of a sound. In this manner, we could refer to that structure when defining *paragoge* (addition at the end

of a word), *epenthesis* (addition within a word), *prothesis* (addition at the beginning of a word), etc.

In example 2 the attribute `@feats` refers to the features related to the sound addition, so that the phenomena “paragoge” and “prothesis” can be characterized simply by adding the features that define them more precisely.

Example 2. “Reusing” features to define phenomena that entail the addition of a sound.

```
<fs xml:id="sound-addition">
 <f name="taxonomy">
  <fs type="linguistic">
   <f name="category">
    <fs type="phonetic">
     <f name="process">
      <symbol value="addition"/>
     </f>
    </fs>
   </f>
  </fs>
 </f>
</fs>
<fs xml:id="paragoge">
 <f name="taxonomy">
  <fs feats="#sound-addition">
   <f name="position">
    <symbol value="end"/>
   </f>
   <f name="constriction">
    <binary value="false"/>
   </f>
  </fs>
 </f>
 <f name="description">
  <string xml:lang="en">Paragoge</string>
 </f>
</fs>
<fs xml:id="prothesis">
 <f name="taxonomy">
  <fs feats="#sound-addition">
   <f name="position">
    <symbol value="start"/>
```

```
    </f>
   </fs>
  </f>
  <f name="description">
   <string xml:lang="en">Prothesis</string>
  </f>
 </fs>
```

21 Furthermore, the feature structure system is so flexible that it allows the combination of heterogeneous types of categorizations using the same methods. This may be especially convenient when working with variant taxonomies and scholarly editions, since different theories and models may require different categories, and it may be of interest to incorporate them simultaneously. See example 3 for a description of dittography, a type of scribal error characterized by a repetition, that could be represented within the same master taxonomy that would include examples 1 and 2.

Example 3. Entry in the taxonomy for a particular scribal error.

```
<fs xml:id="dittography">
 <f name="taxonomy">
  <fs type="error">
   <f name="category">
    <fs type="involuntary">
     <f name="process">
      <symbol value="eye-skip"/>
     </f>
     <f name="result">
      <symbol value="repetition"/>
     </f>
    </fs>
   </f>
  </fs>
 </f>
 <f name="description">
  <string xml:lang="en">Dittography</string>
 </f>
</fs>
```

22 In this case, the outer layers of the definition relate the three examples to one another, although the inner elements used to analyze them further are completely different.

4. Internal Validation of the Taxonomy

23 Because a taxonomy may have many diverse layers of information, it would be easy to lose track of the features, adversely affecting the accuracy of data retrieval. That is why the Feature System Declaration (TEI Consortium 2016, 18.11) is crucial: it is the instrument for declaring all attributes and their possible values, that is, for listing all feature names and feature values, to which a prose description may be added to explain what each represents. This declaration forces the creation of well-documented taxonomies, which provides advantages for the reuse and especially the development of these taxonomies.

24 Additionally, the feature system declaration provides a means for defining constraints, describing what a well-formed and valid feature structure is according to the theory and praxis of the research that is being developed. Alongside the documentary value of this definition, its typed-feature

modeling facilitates the creation of schema constraints. For instance, I process my declaration to further constrict my schema so the feature structure declaration and its actual application are always synchronized and up to date.[3]

Example 4. Excerpt of a feature system declaration.

```
<fsdDecl>
 <fsDecl type="variants.taxonomy">
  <fDecl name="description">
   <fDescr>Prose description of the phenomenon</fDescr>
   <vRange>
    <vNot>
     <string/>
    </vNot>
   </vRange>
  </fDecl>
  <fDecl name="taxonomy">
   <fDescr>Main categorization for type of variant</fDescr>
   <vRange>
    <vAlt>
     <fs type="linguistic"/>
     <fs type="error"/>
     <fs type="material"/>
     <fs type="equipollent"/>
     <fs type="graphic"/>
    </vAlt>
   </vRange>
  </fDecl>
 </fsDecl>
 <fsDecl type="linguistic">
  <fDecl name="category">
   <vRange>
    <vAlt>
     <fs type="phonetic"/>
     <fs type="morphosyntatic"/>
     <fs type="lexical"/>
     <fs type="language-transfer"/>
    </vAlt>
```

```
    </vRange>
   </fDecl>
  </fsDecl>
  <fsDecl type="phonetic">
   <fsDescr>Features to describe phonetic change</fsDescr>
   <fDecl name="process">
    <fDescr>Feature to represent the main type of sound change</fDescr>
    <vRange>
     <vAlt>
      <symbol value="addition"/>
      <symbol value="reduction"/>
      <fs type="alteration"/>
     </vAlt>
    </vRange>
   </fDecl>
   <fDecl name="constriction">
    <fDescr>Vowel or consonant phenomenon</fDescr>
    <vRange>
     <vAlt>
      <binary value="true"/>
      <binary value="false"/>
     </vAlt>
    </vRange>
   </fDecl>
   <!-- Continue with the description of other features -->
  </fsDecl>
 </fsdDecl>
```

25 In the sample presented in example 4, there are two features in the outer layer, the feature `"description"`, whose value cannot be an empty string, and the `"taxonomy"`, which can contain any of the following nested feature structures: `"linguistic"`, `"error"`, `"material"`, `"equipollent"`, or `"graphic"`. Similar declarations are built according to the same model to describe linguistic feature structures and the individual features that define them. When there is no need to go more deeply into the decomposition of a feature structure, and the possible values conform to a limited list, the `<symbol>` element is used to define this controlled vocabulary. In the case of boolean-type values, as in the feature `"constriction"` seen in example 1, the `<binary>` element is declared instead in order to define the constriction.

5. Encoding Textual Variation in TEI

26 As is also often the case elsewhere in the TEI Guidelines, there are several ways to encode variants, as well as alternative methods for linking the apparatus information to the text (see the Critical Apparatus documentation for more information: TEI Consortium 2016, 12). Of the available methods, the parallel segmentation method (TEI Consortium 2016, 12.2.3) seems to be a popular encoding technique for multi-witness editions, in terms of both the specific tools that have been created for this method and the number of projects that apply it.[4] The discussion below explores the integration of a variant taxonomy into an edition that follows this method by inserting an <app> element for each variation unit, that is, in every *locus* in the text where at least two concurrent readings exist (Macé, De Vos, and Geuten 2012, 113).

27 Taxonomies can be formalized as complex modules of structured information, and in the interest of maintaining legibility for human editors, an efficient way to incorporate analytic information into an edition involves the use of stand-off annotation methods (Bański 2010). Stand-off refers to annotation that is not inserted in line. It usually entails the development of the annotation of a primary document in a different file or files from the one that contains the primary textual data. The process of relating the primary document to its annotation involves linking between specific locations of the primary source and the information that describes them, whether through byte offsets, elements, attributes, or other methods (Ide and Romary 2004, 218).

28 In simpler traditions, a semantic correspondence through the use of the attribute @ana, as in the examples below, may also be suitable. Each entry of the taxonomy has an ID that is referred to in an @ana attribute in the edition.

Example 5. Excerpt from a multi-witness edition.

```
<l n="13">
 <app>
  <rdg wit="#A" ana="#np">se<seg>n</seg>pre</rdg>
  <rdg wit="#B" ana="#abb"><choice><abb>semp̃</abb><expan>semp<ex>re</ex></
expan></choice></rdg>
 </app>
 <app>
  <rdg wit="#A" ana="#reg"><seg>ll</seg>e</rdg>
  <rdg wit="#B" ana="#li">lh<seg>i</seg></rdg>
 </app>
 <app>
  <rdg wit="#A" ana="#pal-stem">qui<seg>ge</seg></rdg>
  <rdg wit="#B" ana="#abb"><choice><abb>q̈s</abb><expan>q<ex>ui</ex>s</expan></
choice></rdg>
 </app>
 <app>
  <rdg wit="#A" ana="#gap"/>
  <rdg wit="#B">muj</rdg>
 </app>
 <app>
  <rdg wit="#A" ana="#reg">me<seg>ll</seg>or</rdg>
  <rdg wit="#B" ana="#abb"><choice><abb>melh̃</abb><expan>melh<ex>or</ex></
expan></choice></rdg>
 </app> toda <app>
  <rdg wit="#A">uia</rdg>
  <rdg wit="#B" ana="#y-minim">uya</rdg>
 </app>
</l>
```

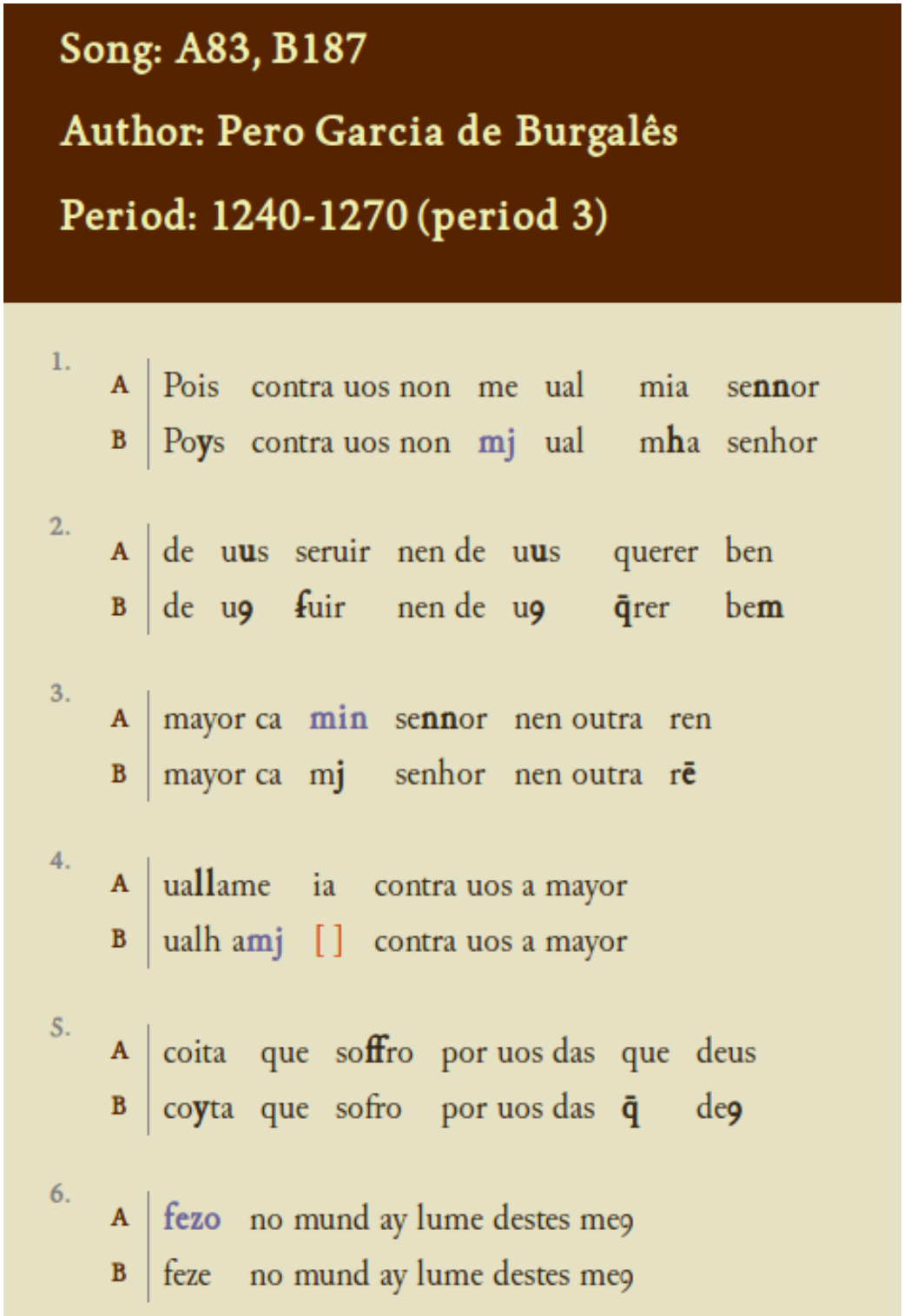
29 Example 5 shows a line of the corpus, encoded according to the TEI parallel segmentation method. The text nodes that are direct children of <l> are common text shared by all witnesses, and an *apparatus* element, <app>, is introduced wherever there are divergences. If there is more than one variant per <app>, the element <seg> encloses the affected characters in order to avoid ambiguities

regarding which part of the token refers to which variant. If there were two or more variants inside the same token, then the `<seg>` element would contain a `@corresp` attribute whose value would be the ID of the variant. When there are additional elements that provide the required semantics for the identification of the characters related to the variant, the use of `<seg>` is avoided (see in example 5 how the variants related to the use of abbreviations are encoded with specific markup which prevents any possible ambiguity). This strategy allows an accurate retrieval of any instance of the phenomena defined in the taxonomy.

6. Additional Analyses

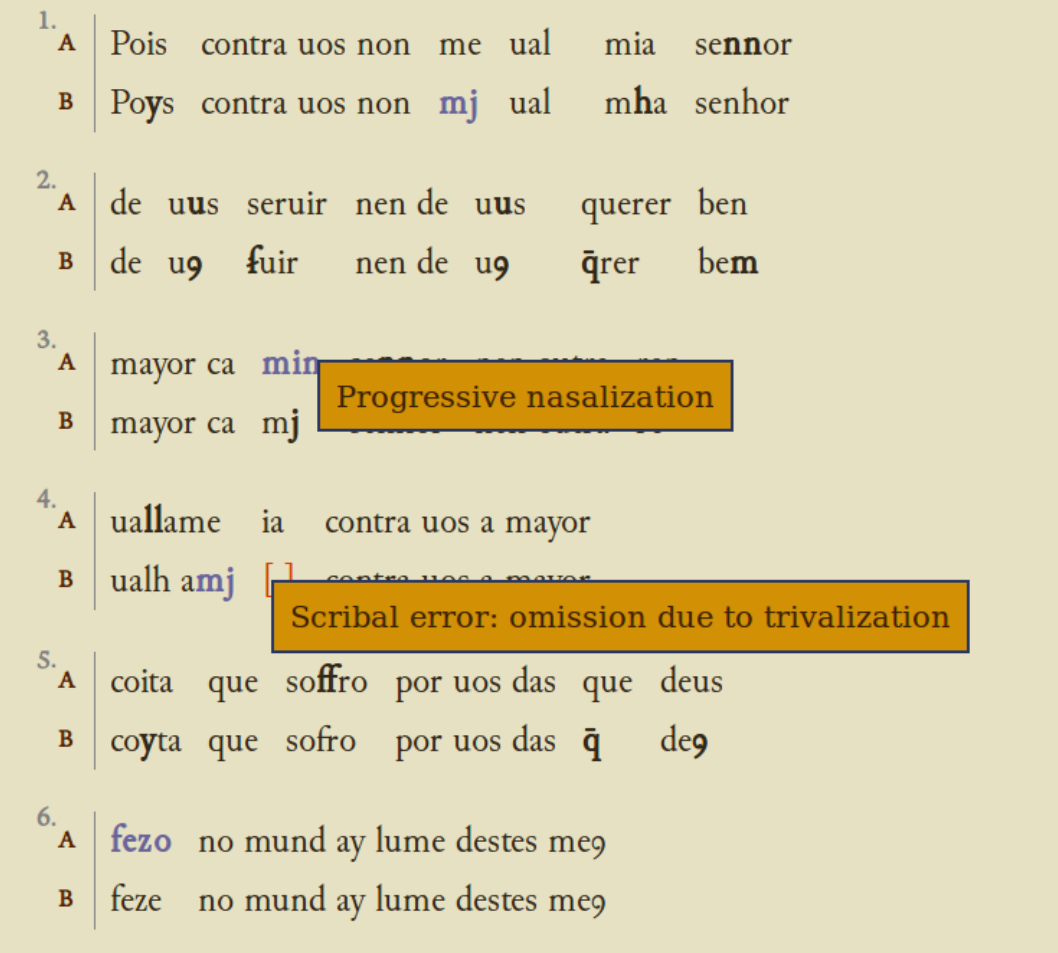
30 One of the functions that the variant taxonomy can fulfill in the publication of the edition is the provision of an accurate description for each textual variant. One way to explore the use of the taxonomy is through enhancing the edition by using visual cues to define the type of variant.

Figure 1. HTML edition sample.



31 The screenshot presented in figure 1 shows two lines of the edition of one of the compositions of my corpus. I have used the higher categories to color-code the variants: graphic variants are bold, linguistic ones are indigo, errors are red, and equipollent readings are underlined. The use of colors is complemented with tooltips, so there is no information that it is only conveyed through color. If we click on any of the variants, we retrieve their more specific descriptions, that is, the contents of `f[@name eq "description"]`, as shown in figure 2

Figure 2. HTML edition sample.



32 In the same way that we access the different hierarchies of the taxonomy to enrich the edition, we can query the textual variants.

33 For instance, we can create a web form following the classification of the different subcategories. This would allow us to explore the frequencies of these variants in the corpus (figure 3). This type of approach makes it possible to study each variation phenomenon by calculating its distribution according to witness and scribe, by period of composition, and, of course, by analyzing all its occurrences in the corpus (figure 4).

Figure 3. HTML web form of variation phenomena.

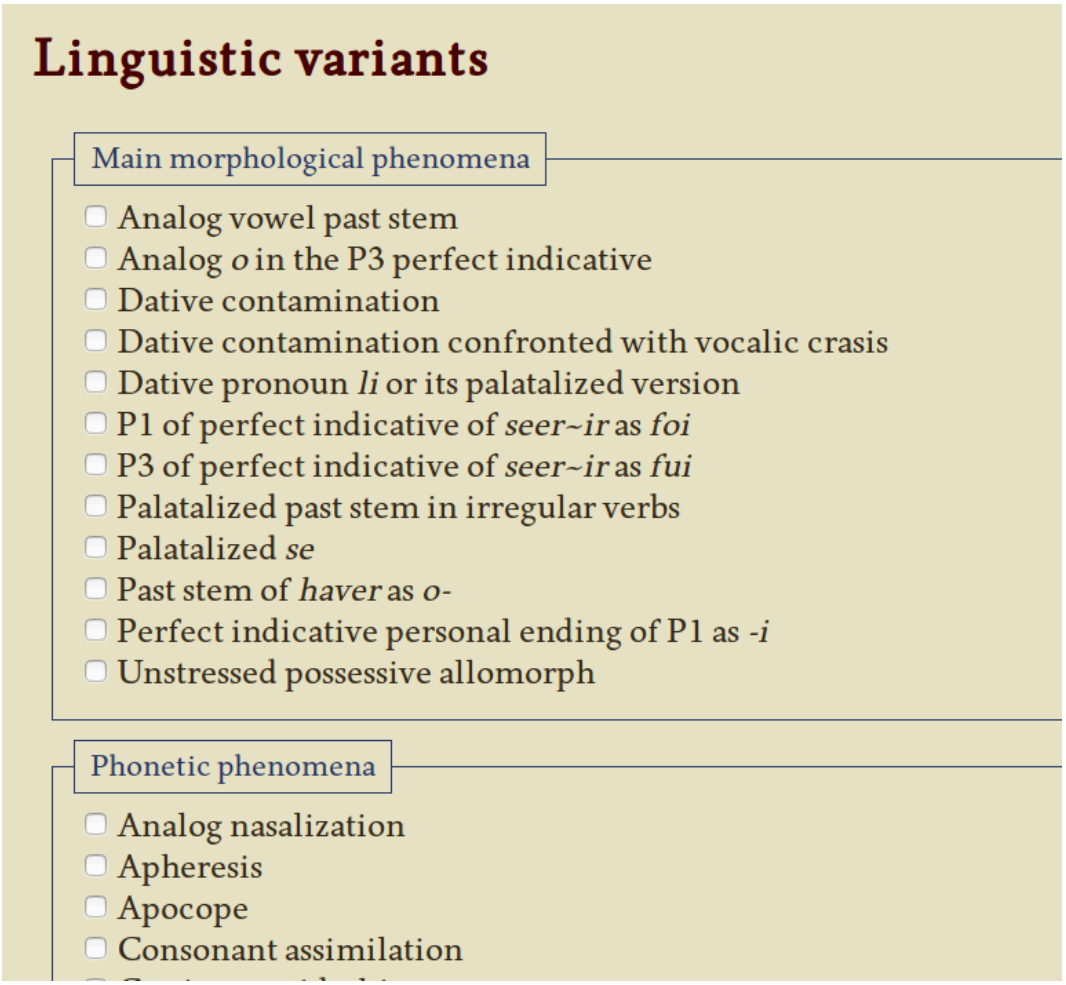
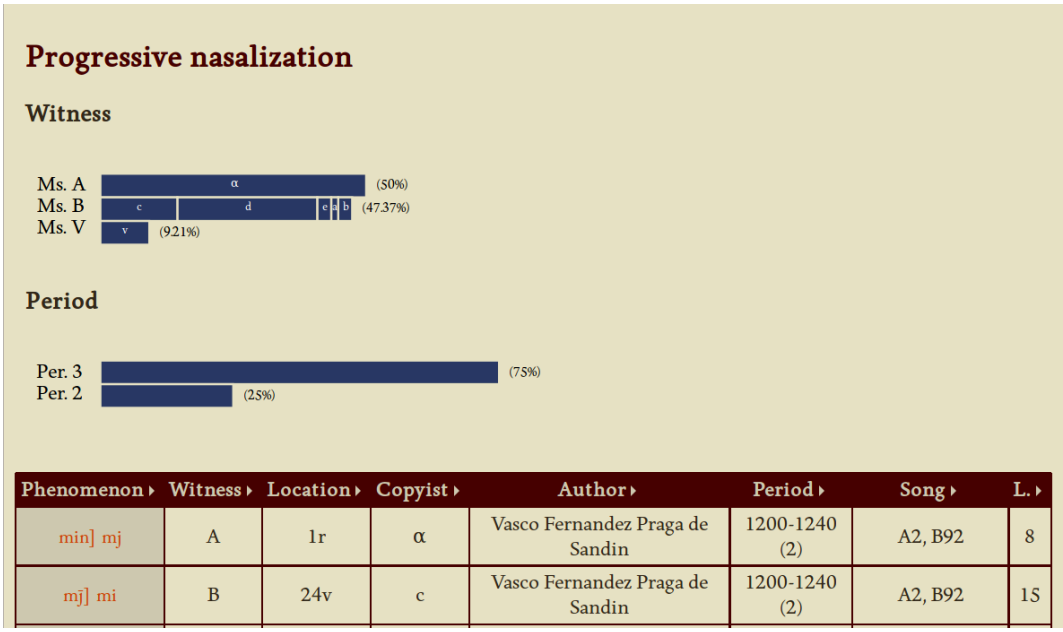


Figure 4. Results after selecting a phenomenon.



Phenomenon	Witness	Location	Copyist	Author	Period	Song	L.
min] mj	A	1r	α	Vasco Fernandez Praga de Sandin	1200-1240 (2)	A2, B92	8
mj] mi	B	24v	c	Vasco Fernandez Praga de Sandin	1200-1240 (2)	A2, B92	15

34 The strong hierarchical organization of feature structures allows us to access efficiently any of the layers of data used in its definition. For example, considering the structure of the taxonomy partially declared in example 4, it might be of interest to retrieve all instances of a linguistic variant related to a phonetic change and compare them numerically to variation related to morphological change.

35 However, the full potential of feature structures becomes clearer with complex queries. Consider, for example, the corpus of syllabic verse compositions mentioned earlier. In such a tradition it would be important to be able to retrieve all lines in which either a linguistic variant or a divergent reading would result in a change in the number of syllables. Such composite queries could determine which linguistic phenomena are likely to represent innovations according to whether they provoke a metrical irregularity, and they could detect a correlation among linguistic phenomena and consequential *emendationes*.

36 Complex queries can be implemented with a small piece of code such as the one presented in example 6. First, I look in the taxonomy for the feature structure IDs in whose definition there are features that would alter the number of syllables: those containing the element `<symbol>` with the attribute values `"addition"` or `"repetition"` (the `$additionPhenomena` variable) or those

with the "reduction" and "omission" values ($reductionPhenomena variable). These include both linguistic variants and scribal errors. Then I look for all lines of the corpus that contain <rdg> elements with at least one reference to each of these two groupings of variant typologies.

Example 6. Query for combining features.

```
declare namespace tei = "http://www.tei-c.org/ns/1.0";
declare variable $lines as node()+ := collection('/db/edition')//tei:div[@type eq
 'poem']//tei:l;
declare variable $features as node()+ := doc('/db/feature-library.xml')//
tei:fvLib/tei:fs;
declare variable $additionPhenomena as xs:string+ := $features[.//tei:symbol/
@value = ('addition', 'repetition')]/concat('#', @xml:id);
declare variable $reductionPhenomena as xs:string+ := $features[.//tei:symbol/
@value = ('reduction', 'omission')]/concat('#', @xml:id);

for $line in $lines[descendant::tei:rdg[some $ana in tokenize(@ana, '\s+')
 satisfies $ana = $reductionPhenomena]]
return $line[descendant::tei:rdg[some $ana in tokenize(@ana, '\s+') satisfies $ana
 = $additionPhenomena]]
```

37 Such queries retrieve occurrences like the one presented in example 5, an authentic example of my corpus in which a linguistic phenomenon might have motivated the scribe to correct the contents of the line in order to regularize the metric pattern. In this line ("Yet, I always loved her more") there is a linguistic phenomenon that creates an extra syllable, the palatalization of the past stem of an irregular verb, which requires a paragogic vowel for pronunciation (*quige* versus *quis*). The witness that does not contain this phenomenon presents extra textual content, the monosyllabic word *muj*, which has an emphatic sense and is therefore omissible without changing the denotative meaning of the line. A plausible hypothesis for explaining this variation is that "quige" is an innovation that motivated a conscious omission of "muj" in order to maintain the correct number of syllables per verse. These nonaccidental omissions or additions are effectively retrieved when the description of linguistic phenomena includes a feature that mentions the addition or reduction of phonemes (as seen in the sample presented in example 4), which enables us to construct queries that look for the co-occurrence of those types of variation with variants related to the textual content.

7. Conclusions

38 Variation is a complex and multifaceted issue. For that reason, a hierarchical model based on the accumulation and nesting of layers of information and able to represent any concepts that depend on categories, subcategories, and even the overlapping of categories is necessary for representing all of these nuances.

39 The examples presented in this paper were modeled based on a specific project and its research questions,[5] but the intention was to present a more general method through a particular application. Nevertheless, the tradition used for exemplification is quite homogeneous and the maximum number of witnesses for the same piece of text is three. This means that a semantic correspondence presented through the use of an attribute in the edition whose value points to the taxonomy might not be suitable for more complex traditions. However, alternative stand-off methods should overcome those limitations. This will be one of the focus points in the future development of a more solid editorial model whose defining feature will be its aptness for descriptive and quantitative analyses of textual variants. Following the distinction made by Jannidis and Flanders (2013), the future work will entail the transformation of an egoistic modeling, designed for a specific research question, to an altruistic one.

40 In spite of its limitations, the core of the methodology presented here might be of interest for other projects. The creation of a variant taxonomy encoded using the feature structures model is a flexible method which brings multiple advantages for textual scholarship. On the one hand, a granular definition of variation phenomena whose information can be embedded later into the edition entails a descriptive model that helps the user browse through the witnesses' readings. On the other hand, it enables quantitative analyses with greater precision and efficiency.

BIBLIOGRAPHY

Bański, Piotr. 2010. "Why TEI Stand-off Annotation Doesn't Quite Work: And Why You Might Want to Use It Nevertheless." In *Proceedings of Balisage: The Markup Conference 2010*. Balisage Series on Markup Technologies, vol. 5. doi:10.4242/BalisageVol5.Banski01.

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

Chiesa, Paolo. 2002. *Elementi di critica testuale*. Bologna: Pàtron Editore.

Colwell, E. C., and E. W. Tune. 1964. "Variant Readings: Classification and Use." *Journal of Biblical Literature* 83 (3): 253–61. doi:10.2307/3264283.

Ide, Nancy, and Laurent Romary. 2004. "International Standard for a Linguistic Annotation Framework." *Natural Language Engineering* 10 (3–4): 211–25. doi:10.1017/S135132490400350X. https://doi.org/10.1017/S135132490400350X.

Italia, Paola, Fabio Vitali, and Angelo Di Iorio. 2015. "Variants and Versioning Between Textual Bibliography and Computer Science." In *Proceedings of the Third AIUCD Annual Conference on Humanities and Their Methods in the Digital Ecosystem* (AIUCD '14). New York: ACM. doi:10.1145/2802612.2802614.

Jakobson, Roman. 1971. *Selected Writings*. Vol. 1, *Phonological Studies*. The Hague: Mouton.

Jannidis, Fotis, and Julia Flanders. 2013. "A Concept of Data Modeling for the Humanities." In *Digital Humanities 2013: Conference Abstracts*, 237–39. Lincoln, NE: Center for Digital Research in the Humanities. http://dh2013.unl.edu/abstracts/ab-313.html.

Maas, Paul. 1958. *Textual Criticism*. Oxford: Clarendon Press.

Macé, Caroline, Ilse De Vos, and Koen Geuten. 2012. "Comparing Stemmatological and Phylogenetic Methods to Understand the Transmission History of the 'Florilegium Coislinianum.'" In *Ars Edendi* Lecture Series, edited by Alessandra Bucossi and Erika Kihlman, 2:107–29. Studia Latina Stockholmiensia 58. Stockholm: Stockholm University Library. urn:nbn:se:su:diva-79673; PDF available at http://www.diva-portal.org/smash/get/diva2:551286/FULLTEXT01.pdf.

Mink, Gerd. 2000. "Editing and Genealogical Studies: The New Testament." *Literary and Linguistic Computing* 15 (1): 51–56. doi:10.1093/llc/15.1.51.

Pose, Javier, Patrice Lopez, and Laurent Romary. 2014. "A Generic Formalism for Encoding Stand-off Annotations in TEI." https://hal.inria.fr/hal-01061548, version 1.

Romary, Laurent. 2015. "Standards for Language Resources in ISO — Looking Back at 13 Fruitful Years." https://arxiv.org/abs/1510.07851.

Segre, Cesare. 1995. "Critique des variantes et critique génétique." *Genesis (Manuscrits-Recherche-Invention)* 7: 29–45. doi:10.3406/item.1995.994.

TEI Consortium. 2016. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 3.1.0. Last updated December 15. N.p.: TEI Consortium. https://www.tei-c.org/Vault/P5/3.1.0/doc/tei-p5-doc/en/html/.

Trovato, Paolo. 2014. *Everything You Always Wanted to Know about Lachmann's Method. A Non-Standard Handbook of Genealogical Textual Criticism in the Age of Post-Structuralism, Cladistics, and Copy-Text*. Padova: libreriauniversitaria.it.

Witt, Andreas, and Jens Stegmann. 2009. "TEI Feature Structures as a Representation Format for Multiple Annotation and Generic XML Documents." In *Proceedings of Balisage: The Markup Conference 2009*. Balisage Series on Markup Technologies, vol. 3. doi:10.4242/BalisageVol3.Stegmann01.

NOTES

1 This statement is especially significant when dealing with corpora that have been compiled over a long period of time. As is clearly explained in the introduction to the Helsinki Corpus that Irma Taavitsainen and Päivi Pahta prepared for the Corpus Resource Database (CoRD) ("Placing the Helsinki Corpus Middle English Section Introduction into Context," http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/meintro.html): "The idea of basing corpus texts directly on manuscript sources has been presented more recently ... The principles of preparing manuscript texts for print have undergone changes during the history of editing ...."

2 For examples of different variant classifications, see Colwell and Tune (1964) and Italia, Vitali, and Di Iorio (2015).

3 I use XSLT to process the feature structure declaration in order to create all required Schematron rules that will constrict the feature library accordingly. I am currently working on creating a more generic validator (see my Github repository, https://github.com/HelenaSabel/FS-Validator).

4 Tools include Versioning Machine, CollateX (both the Java and Python versions), and Juxta. For representative projects using the parallel segmentation method see Satire in Circulation: James editions Russell Lowell's Letter from a volunteer in Saltillo, Walden: A Fluid-Text Edition, or Digital Mitford: The Mary Russel Mitford Archive, to name a few.

5 The working hypothesis of the aforementioned project entailed a quantitative analysis of linguistic variants to provide evidence that the concentration of certain archaisms on the one hand and innovations on the other will depict the different stages in the compilation of textual materials. The objects of study were medieval songbooks and the transmission of the particular tradition under analysis is based on the reuniting of smaller songbooks that were disseminated independently before creating the preserved compilations.

AUTHOR

HELENA BERMÚDEZ SABEL

Helena Bermúdez Sabel is a postdoctoral researcher at the Université de Lausanne. Currently, she is working for the SNFS-funded project *A world of possibilities. Modal pathways on the extra-long period of time: the diachrony of modality in the Latin language* which aims at reconstructing the evolution of modal meanings from the prehistory of the Latin language up to the seventh century CE. In this project, Helena supervises the technical aspects of the annotation workflow, including the automation of corpus pre-processing, annotation and publication.