Lightweight Grammatical Annotation in the TEI: New Perspectives

Piotr Bański[1], Susanne Haaf[2], Martin Mueller[3]
[1]Institute for the German Language, Mannheim (Germany)
[2]Berlin-Brandenburg Academy of Sciences and Humanities, Berlin (Germany)
[3]Northwestern University, Chicago (USA)
[1]banski@ids-mannheim.de, [2]haaf@bbaw.de, [3]martinmueller@northwestern.edu

Abstract

In mid-2017, as part of our activities within the TEI Special Interest Group for Linguists (LingSIG), we submitted to the TEI Technical Council a proposal for a new attribute class that would gather attributes facilitating simple token-level linguistic annotation. With this proposal, we addressed community feedback complaining about the lack of a specific tagset for lightweight linguistic annotation within the TEI. Apart from `@lemma` and `@lemmaRef`, up till now TEI encoders could only resort to using the generic attribute `@ana` for inline linguistic annotation, or to the quite complex system of feature structures for robust linguistic annotation, the latter requiring relatively complex processing even for the most basic types of linguistic features. As a result, there now exists a small set of basic descriptive devices which have been made available at the cost of only very small changes to the TEI tagset. The merit of a predefined TEI tagset for lightweight linguistic annotation is the homogeneity of tagging and thus better interoperability of simple linguistic resources encoded in the TEI. The present paper introduces the new attributes, makes a case for one more addition, and presents the advantages of the new system over the legacy TEI solutions.

Keywords: linguistic annotation, lightweight annotation, TEI, TEI LingSIG

1. Introduction

In July 2017, as part of our activities within the TEI Special Interest Group for Linguists (LingSIG), we submitted to the TEI Technical Council a proposal for the definition of a new attribute class that gathers token-level attributes facilitating simple linguistic annotation. With this proposal, we addressed repeated requests from various corpus projects to facilitate combined annotation of customary TEI text structures and basic token-level linguistic features. Most of our proposed modifications have been accepted and are part of the new release of the TEI Guidelines (3.3.0, published in January 2018).[1] The present paper is therefore partially a report of success and a review of new features that the encoder has at her disposal, and in part, it provides arguments for extending the descriptive power of the Guidelines even further, towards supporting a complete chain of token-level grammatical analysis in various kinds of text collections.

In the out-of-the-box TEI markup, it is now possible to use the following attributes of the elements `<w>` (« word ») and `<pc>` (« punctuation character »):

- `@pos` (for part-of-speech)
- `@msd` (for morpho-syntactic description)
- `@join` (to signal string concatenation with a neighbouring element)

These attributes – together with `@lemma` and `@lemmaRef`, previously defined inside `<w>` – are now encapsulated in a new attribute class, *att.linguistic*, supplying means crucial for basic grammatical annotation at the token level, i.e., for lightweight grammatical annotation.[2]

In the sections to follow, we provide a description of the new attributes and a discussion of the alternatives that have been in use up till now. A further, argumentative part of the present paper focuses on an additional attribute that facilitates the encoding of historical corpora and literary collections, namely on `@norm`, encoding normalized/ regularized forms.

2. TEI and the TEI Guidelines

« TEI » stands for « Text Encoding Initiative », a consortium of institutions and individuals aiming at developing guidelines for consistent and explicit encoding of a wide array of textual types. The TEI Guidelines are freely available at http://www.tei-c.org/, in the form of prose, documented schemas, and ready-to-use customizations, together with various tools. This section provides a brief overview of the TEI Guidelines essential for contextualizing the rest of the paper.

The TEI Guidelines are encoded in, and customized by, a TEI-based specialized literate markup language called ODD.[3] ODD allows for the definition of TEI-specific constructs (modules, element classes, attribute classes) which can be combined in various ways in order to form descriptive apparatus for a variety of phenomena encountered in Digital Humanities' research. ODD combines such definitions with potentially very extensive documentation (the most extreme form of which is found in the multi-chapter prose of the Guidelines themselves). Apart from tools that can assist the text modeller by manipulating ODD documents in order to produce customized schemas (the most well-known such tool is Roma), encoders may use so-called TEI Extensions which are essentially out-of-the-box customizations designed for particular research areas.

[1] The changes were merged in a pull request which also references detailed discussion, see https://github.com/TEIC/TEI/pull/1671 and issue #1670.

[2] We enclose element names in angle brackets and prepend attribute names with a '@'.

[3] See http://www.tei-c.org/Guidelines/Customization/odds.xml and https://wiki.tei-c.org/index.php/ODD.

Crucially for the proposal presented here, the TEI data model includes constructs known as element classes and attribute classes. Element classes are sets of TEI elements that are found in similar structural contexts. Attribute classes group attributes that have something in common in terms of features modelled in a particular domain. When an element is a member of an attribute class, it can use all the attributes found inside that class.

Our focus here is on the elements <w> and <pc> that are members of the model.segLike element class and are specialized for the description of linguistic tokens and punctuation characters, respectively. Each of these elements is a member of several attribute classes. An element can also, in principle, define attributes that are specific to it and ideally not needed anywhere else – that is currently the case of <pc>, and it was the case of <w>, which used to define @lemma and @lemmaRef before the changes described here came into effect.[4]

The present paper is centred around the newly added attribute class *att.linguistic*. Postulating a new attribute class is a theoretical statement as well as a practical move. From the theoretical point of view, attributes contained in a single class should have something in common: in our case, they are necessary for a reasonably flexible description of basic grammatical features of tokens. From the practical perspective, defining a single class of attributes that function together means that the set of properties that it provides can be made available to other elements, if needed, via ODD customization.

3. Grammatical Annotation in the TEI prior to the *att.linguistic* class

Enrichment of TEI-annotated text with even the most basic grammatical information (part-of-speech and morphosyntactic information in addition to lemma identification) drastically expands search options, options for sorting the results, or for investigating author-specific traits. It also facilitates the combined analysis of token-based linguistic information with specific TEI structures, e.g. investigation of adjective usage in headlines, etc. (see e.g. Schöch 2016, Haaf 2016). Over the years, many robust structural solutions have been suggested, but no single standardized approach for lightweight annotation at the token level has emerged. This section looks briefly at the solutions suggested so far and shows why some of them are not optimal.

3.1 Hierarchical solutions

The TEI possesses tools for complex description of linguistic structures by means of element hierarchies, among others by exploiting the extremely powerful mechanism of feature structures, defined in a joint ISO-TEI standard (ISO 24610-1, TEI Consortium, 2018, Ch. 18 ; see fig. 1 below for an illustration). Indeed, as shown by Stegmann and Witt (2009), it is conceivable, even if not practical, to model an entire linguistic corpus as a complex feature matrix. Typically, feature structures are used for modelling local bundles of features, grammatical and other, that would otherwise not fit into the format prescribed by the TEI. They are a very handy descriptive device in all formats that use the stand-off approach, where annotations are not part of the (sub)document containing annotated text.[5]

An example of a feature structure that at the same time illustrates much of the functionality that *att.linguistic* now provides, comes from the National Corpus of Polish[6] and represents a set of potential interpretations of the adjectival form *kategoryczne* (« categorical » ; fig. 1). In the example, the feature « base » has the lemma as its value, and « ctag » encodes the part-of-speech. The feature « msd » lists all possible morphological interpretations of the token (the first option encodes the features : plural, nominative, animate masculine, and positive).[7] Another feature, not shown in the example, is used to point at the value(s) disambiguated in the given morphosyntactic context.

```
<f name="interps">
 <fs type="lex">
  <f name="base">
   <string>kategoryczny</string>
  </f>
  <f name="ctag">
   <symbol value="adj"/>
  </f>
  <f name="msd">
   <vAlt>
    <symbol value="pl:nom:m2:pos"/>
    <symbol value="pl:nom:m3:pos"/>
    <symbol value="pl:nom:f:pos"/>
    <symbol value="sg:nom:n:pos"/>
    <symbol value="pl:nom:n:pos"/>
    <symbol value="pl:acc:m2:pos"/>
    <symbol value="pl:acc:m3:pos"/>
    <symbol value="pl:acc:f:pos"/>
    <symbol value="sg:acc:n:pos"/>
    <symbol value="pl:acc:n:pos"/>
    <symbol value="pl:voc:m2:pos"/>
    <symbol value="pl:voc:m3:pos"/>
    <symbol value="pl:voc:f:pos"/>
    <symbol value="sg:voc:n:pos"/>
    <symbol value="pl:voc:n:pos"/>
   </vAlt>
  </f>
 </fs>
</f>
```

Figure 1. National Corpus of Polish, file « ann_morphosyntax.xml », with ID attributes stripped off for the sake of conciseness

Another way to use feature structures is by defining feature libraries and feature-value libraries, where individual elements can be referenced by the attribute

[4] For <w> see the penultimate release: http://www.tei-c.org/Vault/P5/3.2.0/doc/tei-p5-doc/en/html/ref-w.html; for <pc> see: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-pc.html.

[5] For references and remarks on TEI standoff techniques, see a.o. Bański, 2010, Pose et al., 2014, and Bański et al., 2016, as well as https://github.com/laurentromary/stdfSpec.

[6] See http://nkjp.pl/ for an entry point and information, and http://nlp.ipipan.waw.pl/TEI4NKJP/ for ODD and examples.

[7] See the tagset documentation for other categories and values: http://nkjp.pl/poliqarp/help/ense2.html.

@ana, as in fig. 2, which shows an example copied from Budin et al., 2012:11.

```
<fLib>
 <f xml:id="pos.verb" name="pos">
  <symbol value="verb"/> </f>
 ...
 <f xml:id="tns.pres" name="tense">
  <symbol value="present"/> </f>
 ...
 <f xml:id="mood.ind" name="mood">
  <symbol value="indicative"/> </f>
 ...
 <f xml:id="num.pl" name="number">
  <symbol value="plural"/> </f>
 ...
 <f xml:id="pers.1" name="person">
  <symbol value="1"/> </f>
 ...</fLib>

<fvLib> ...
 <fs   xml:id="v_pres_ind_sg_p2"
       name="v_pres_ind_sg_p2"
       feats="#pos.verb #tns.pres #mood.ind
              #num.pl #pers.2">
 ...</fvLib>

<form  type="inflected"
       ana="#v_pres_ind_pl_p1
              #v_pres_ind_pl_p3 ">
 <orth>gehen</orth>
</form>
```

Figure 2. <fLib> contains simple attribute-value pairings, while <fvLib> can be used to create complex values for re-use (multiple references to single pairings are grouped into bundles). Both simple and complex features can then be referenced, from any language resource, by means of the @ana attribute. (Copied from Budin et al., 2012:11)

A decision to use such complex hierarchical element-based devices for grammatical description indicates a commitment in terms of various resources: manpower, time, finances needed to create and/or customize and maintain specialized tools capable of interpreting such robust structures. In practice, this kind of commitment is not always possible or needed. A tokenized corpus where element hierarchy is relatively simple and where grammatical features are bundled inside word-sized textual segments, is able to support or reject many linguistic hypotheses at a much lower cost than a robust resource would incur. Similar observations are true of an average case of literary encoding, to which grammatical information gets added for the purpose of enhancing searches or for basic measurements – it can be added as a separate document, i.e. in a stand-off manner, but in many cases it is much simpler and cheaper to add the relevant attributes to the individual <w>-sized segments.

3.2 Solutions for lightweight annotation before TEI version 3.3.0

Before the initiative described here, for the purpose of encoding the results of simple grammatical analysis, the TEI encoder had to resort to non-standardized devices, essentially to using semantically unspecified attributes to carry linguistic description. The primary candidates in this context were @ana, @corresp and @type, where:

- @ana: "indicates one or more elements containing interpretations of the element on which the @ana attribute appears" (TEI 2018: att.global.analytic)
- @corresp: "points to elements that correspond to the current element in some way" (TEI 2018: att.global.linking)
- @type: "characterizes the element in some sense, using any convenient classification scheme or typology" (TEI 2018: att.typed)

Of this attribute set, @ana and @corresp are pointer-based, while @type can hold sequences of whitespace-delimited tokens. Fig. 3 illustrates the use of @ana with an actual example from the Guidelines.

```
<s>
 <w ana="#AT0">The</w>
 <w ana="#NN1">victim</w>
 <w ana="#POS">'s</w>
 <w ana="#NN2">friends</w>
 <w ana="#VVD">told</w>
 <w ana="#NN2">police</w>
 <w ana="#CJT">that</w>
 [...]
<s>
```

Figure 3. Partial example of using @ana advocated by the Guidelines (TEI 2018:17.4)

Our main arguments against the use of the above-mentioned attributes for lightweight markup are based on the notions of simplicity and practicality. Firstly, a large part of our target group, namely corpus linguists and creators of language resources, need a straightforward way to serialize the output of analysis tools, using well-established labels, concepts, and datatypes. An approach using @ana or @corresp would essentially involve creating pseudo-URIs out of the labels produced by morphological analyzers, only to pre-process those URIs for querying and visualisation in order to convert them back to simple labels.

The other major part of our target group, namely the creators and curators of resources for other disciplines (e.g., literary and historical text collections), frequently use any of the three attributes, in any combination, for the purpose of domain-specific text analysis. When these resources become subject to enrichment with linguistic markup, there will be no way to guarantee a uniform choice of containers for grammatical information (or, in the case of @ana and @corresp with added pseudo-URIs, a uniform structuring of values), unless the *att.linguistic* class can be used for this purpose. Reserving some of the generic attributes for linguistic purposes

would mean (a) removing them from the general pool of attributes available for non-linguistic uses and (b) excluding some of the legacy non-linguistic resources that already use those attributes.

On the basis of the above considerations and being aware that identification of the lemma, the POS and the morphosyntactic features are the most basic requirements of linguistic or linguistically-informed analysis at any level of processing, we decided to propose analytical attributes specifically fitted for the linguistic domain and thus to separate linguistic token-based annotation from that used in other domains.

4. Description of the *att.linguistic* class

One notorious point of criticism concerning the TEI is that it 'allows for too much'. This is apparently based on an expectation that there should be a single way to encode any given kind of textual phenomena. However, anyone with moderate awareness of the richness of modern day Digital Humanities will know that there is definitely no single way to approach the different kinds of data, information needs, visualisation requirements and the varying foci of interest of Digital Humanities' scholars. The TEI is a toolkit from which a researcher needs to define and document a particular customization, given the plethora of options on offer. For a "tinkerer", the TEI is a nearly endless collection of parts from which numerous schemas can be created. Not every researcher wants to be a tinkerer, however – some would prefer to be end-users of predesigned solutions and to have at their disposal a ready-made standardized format, which only needs to be filled in with text and annotated values. The solution described here adds low-level devices from which a tinkerer can choose for the purpose of enriching already existing schemas but which at the same time can straightforwardly be used by a corpus linguist planning to create a new digital resource consisting of only crudely structured tokenized text with the basic linguistic features – which now have clearly labelled containers in the form of *att.linguistic* attributes.

An example applying the crucial *att.linguistic* components follows below.

```
<s>
 <w pos="PPER" msd="pers:subst:p1:nom:pl"
       lemma="wir">Wir</w>
 <w pos="VVFIN" msd="p1:pl:pres:ind"
       lemma="fahren">fahren</w>
 <w pos="APPR" msd="--" lemma="in">in</w>
 <w pos="ART" msd="def:acc:sing:masc"
       lemma="d">den</w>
 <w pos="NN" msd="acc:sing:masc"
       lemma="Urlaub">Urlaub</w>
  <pc pos="$." msd="--" lemma=".">.</pc>
</s>
```

Figure 4. The results of an analysis of the sentence « Wir fahren in den Urlaub » encoded by means of *att.linguistic*. Source of analysis: WebLicht (2018)

The example above contains the following attributes:

- `@pos`: « (part of speech) indicates the part of speech assigned to a token, usually according to some official reference vocabulary (e.g. for German: STTS, for English: CLAWS, for Polish: NKJP, etc.) » (TEI 2018 : att.linguistic)
- `@msd`: « (morphosyntactic description) supplies morphosyntactic information for a token, usually according to some official reference vocabulary (e.g. for German: STTS-large tagset, for Polish, NKJP) »[8] (TEI 2018 : att.linguistic)
- `@lemma`: « provides a lemma (base form) for the word, typically uninflected and serving both as an identifier (e.g. in dictionary contexts, as a headword), and as a basis for potential inflections. » (TEI 2018 : att.linguistic)

Note that due to various compromises that have to be made between linguistic description and technological efficiency, it is not unnatural to expect projects to use only one of `@pos` and `@msd` for storing complex information, or to use them redundantly, e.g. with `@pos` containing part-of-speech symbols extracted from composite morphosyntactic labels stored inside `@msd`. It often happens, especially in languages with impoverished inflection, that morphosyntactic categories are merged into parts of speech (this is partially responsible for the difference between, e.g., CLAWS-5 and CLAWS-8). It is expected that each project will document the particular grammatical annotation practices in the corpus header.

In order to illustrate the next member of *att.linguistic*, we repeat the example sentence from fig. 4 in a somewhat different arrangement, and with the attributes removed, for the sake of clarity (fig. 5).

```
<s><w>Wir</w> <w>fahren</w> <w>in</w>
<w>den</w> <w>Urlaub</w><pc>.</pc></s>
```

Figure 5. Tagging of example sentence from fig. 4 as inline representation

If we contrast the inline representation in fig. 5 with the sequential representation in fig. 4, it becomes clear that the example in fig. 5 provides more information, because it uses whitespace as additional typographical markup. The representation in fig. 4 only lists the tokens according to their order in the sentence, but loses the information on the lack (or the presence) of the neighbouring whitespace. In order to preserve this kind of information, the `@join` attribute should be used. With this attribute, the final fragment of the sequential example looks as illustrated in fig. 6.

An issue may arise concerning the redundancy of marking the absence of whitespace on two neighbouring elements. From the top-down, global perspective, it is indeed redundant. From the bottom-up, 'streamable' perspective, it is not redundant, and we assume that decisions on which stance to adopt are going to be project-specific. The TEI

[8] For a feature description system designed as (pragmatically) universal for use with Universal Dependencies, see http://universaldependencies.org/u/feat/index.html; for the corresponding system of parts of speech see http://universaldependencies.org/u/pos/index.html.

provides support for the general, redundant, 'streamable' case, of which project-specific decisions can be subsets.

```
<s>
    <w>Wir</w>
    <w>fahren</w>
    <w>in</w>
    <w>den</w>
    <w join="right">Urlaub</w>
    <pc join="left">.</pc>
</s>
```

Figure 6. Example sentence from fig. 4 in the sequential arrangement, with redundant (=streamable) use of `@join`.

The last attribute of att.linguistic is `@lemmaRef`, previously defined directly inside `<w>`. It provides a pointer to a definition of the lemma for the word, e.g. in an online lexicon, as illustrated in fig. 7 (copied from the Guidelines).

```
<w type="verb" lemma="hit"
lemmaRef="http://www.example.com/lexicon/hi
tvb.xml">hitt<m type="suffix">ing</m></w>
```

Figure 7 : Example for the usage of `@lemmaRef`

The current documentation for the class may be accessed at http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-att.linguistic.html.

5. Normalization of forms

In corpora of historical texts with non-standardized spelling, regularization is a very frequent matter which may be applied not only to words but also to punctuation characters. For corpus queries, the normalization of historical spellings may be useful for search purposes because users do not have to think of possible spellings of a search term. For further linguistic analysis, the normalization of writings may be a significant first step towards providing homogeneous input data for applications based on lexica with standardized spelling (e.g., lemmatization and POS analysis usually expects modernized spellings, cf. Jurish 2012[2008]:13). The TEI has at its disposal the powerful but also a somewhat costly descriptive mechanism of `choice/reg|orig` for the annotation of regularization, whereby the element `<choice>` contains the element `<orig>` with the original version of the text, and the element `<reg>` with the normalized version. For the purpose of lightweight linguistic description and for the sake of coherence with the already adopted proposals, we suggest an addition of an alternative device – an attribute to store the normalized equivalent of the text content of `<w>` or `<pc>`. That attribute is already part of the TEI repertoire, but it is defined by another attribute class, namely *att.lexicographic*:[9]

[9] In contrast to research areas dealing with strictly numeric data of various sorts, the langage-resource community uses the terms « normalization », « standardization », « regularization » (and often also « modernization ») to a large extent interchangeably, with nuances getting teased out only at the level of specific project guidelines and only when needed. This is why there is nothing untoward in suggesting the use of `@norm` for the functionality otherwise claimed for the element named `<reg>` (defined in the Guidelines as containing « a reading which has been regularized or normalized in some sense »).

`@norm` : « (normalized) gives a normalized form of information given by the source text in a non-normalized form. » (TEI 2018 : att.lexicographic) ; status: optional; datatype: teidata.text[10]

We propose that this attribute should also be available within the class *att.linguistic*. In other words, we do not postulate an introduction of a new device, but merely an extension of the structural context in which an existing TEI attribute may be used, keeping its definition and data type intact.

The `@norm` attribute would complete the set of attributes for token-based linguistic annotation. As stated above, (automatic) normalization is a crucial basic step for further linguistic analysis on the token level. In this sense, tokenization, orthographic normalization, lemmatization, POS tagging and morphosyntactic analysis form a sequence of analytic steps based on one another and thus connected in terms of an analysis chain. The initial point of this analysis is the textual content of the `<w>` element, i.e. a token of the historical source text. The `<w>` element, together with its *att.linguistic* attributes, would then form a single coherent unit encapsulating token-level linguistic information.

By using `choice/reg|orig`, the initial step of the above-mentioned information chain is moved out of the immediate context of the `<w>` element and into another subset of TEI elements. Regarding consistency, it doesn't seem appropriate that some linguistic analysis results for tokens lead to further embedding of the source text whereas others do not. Furthermore, such encoding adds significantly to the complexity of annotation itself. Linguistic annotation becomes mixed with customary TEI encoding, e.g. with the annotation of highlighting (`<hi>`; see fig. 8 and 9), of erroneous text (`choice/sic|corr`), or even with text interrupting the running text (`<fw>`; see fig. 10). From the perspective of processing and post-processing, this mixture of approaches necessitates efforts (e.g. by increasing the possibility of exceptions, or by simply enforcing the usage of different routines to be able to add and extract similar-level information to and from the text) that seem avoidable by allowing for the homogeneity of linguistic markup.

In contrast, the TEI tagset can easily tolerate an addition of an encoding variant that provides a localized alternative to existing tagging solutions. The TEI has not been maintained as a tagset without ambiguities but has rather been created with the motive of suiting as many communities and project necessities as possible:

> « Because of its roots in the humanities research community, the TEI scheme is driven by its original goal of serving the needs of research, and is therefore committed to providing a maximum of comprehensibility, flexibility, and extensibility. [...] This has led to a number of important design decisions, such as: [...] alternative encodings for the same textual features » (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/AB.html#ABTEI2)

[10] See http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-att.lexicographic.html.

The ODD mechanism was created to allow projects to reduce the TEI tagset to a subset adjusted for the respective contexts of its usage:

« Because the TEI Guidelines must cover such a broad domain and user community, it is essential that they be customizable: [...] Customization is a central aspect of TEI usage and the Guidelines are designed with customization in mind. » (http://www.tei-c.org/Guidelines/Customization/)

```
<choice>
 <orig>
  <w lemma="wohlstilisierend" pos="ADJA">
Wohl-<hi rendition="#aq">ſtylifi</hi>rende
  </w>
 </orig>
 <reg>
  <w lemma="wohlstilisierend" pos="ADJA">
Wohl<hi rendition="#aq">stylisie</hi>rende
    </w>
  </reg>
</choice>
```

Figure 8: From Marperger (1717: 655), tagging with `<choice>`

```
<w lemma="wohlstilisierend" pos="ADJA"
  reg="Wohlstilisierende">
Wohl-<hi rendition="#aq">ſtylifi</hi>rende
</w>
```

Figure 9: From Marperger (1717: 655), tagging with `@reg`

```
<w>Flecken</w>
<w>oder</w>
<w norm="Dorf">Dorff</w>
<w norm="desselbigen">deſ-<lb/>
  <fw place="bottom" type="catch">
    ſelbi-</fw>
  <pb facs="#f0672" n="656"/>
  <fw place="top" type="header">
    <hi rendition="#b">Von der</hi>
    <hi rendition="#aq #i">Præſtan</hi>
    <hi rendition="#b">tz
    und Vortreflichkeit</hi>
  </fw><lb/>
  ſelbigen</w>
<w>Landes</w>
```

Figure 10: Fragment from: Marperger (1717: 671); tagging of a token interrupted by page break with `@norm` attribute

It is therefore common that the TEI offers several possible encodings for similar phenomena. Hence, the recurring strand of thought in the discussion of any modifications or enrichment of the TEI, that the system should prevent the encoder from 'making mistakes' in choosing the wrong one out of several tagging solutions, does not fit the constitutive design of the ODD-based TEI. The responsibility of ensuring the 'right' markup within a project is in the hands of the encoder and/or the tools and validation scenarios, which may even be independent of the TEI-defined schemas. From our point of view, the multitude of potential research foci, tools for linguistic analysis and visualisation, and language-specific constraints, necessitate a variety of approaches to tagging for projects to choose from.

We therefore stress that adding `@norm` to *att.linguistic* is not meant to be a replacement for the `choice/reg|orig` system, but rather a localized alternative, to be used where feasible, and primarily for the purpose of adding basic token-level linguistic information.

In terms of an actual implementation, we propose to extract `@norm` from the *att.lexicographic* class into a separate class (call it e.g. *att.normalize*), of which both *att.lexicographic* and *att.linguistic* will then become members. This would create an inheritance hierarchy that would make it possible to avoid duplication of the attribute definition.

6. Application of the Format

At Northwestern University, Philip R. Burns and Martin Mueller have worked on a project to apply simple linguistic annotation and normalization to the not quite two billion words in 60,000 English texts before 1700 and American texts before 1800 and originally transcribed by the Text Creation Partnership (TCP). They have used `@lemma` and `@pos` attributes, as well as a `@reg` attribute that is functionally equivalent to the proposed `@norm` and could be easily replaced by it. This is a coarse-grained and large-scale enterprise where ease of processing becomes an important concern. Things become a lot simpler if all relevant properties of a given 'word' or lexical item can be contained within the element (`<w>` or `<pc>`) that encloses it. It also helps if each property has a readily understood name (`@pos`, `@norm`). If these properties are isolated within an element you can ignore them: they are a sideshow that does not complicate the hierarchical XML structure. You can also extract them more readily: for any analysis of the 'bag of words' type each word comes in a bag that contains the available data about it.

Approximately 7,000 of those texts are currently available at https://texts.earlyprint.org/, to be joined later in 2018 by another 18,000 texts currently in the public domain. These texts exist in an environment that supports collaborative curation. The adoption of the `@reg` (or `@norm`) attribute has made it significantly easier and cheaper to maintain a corpus that is subject to iterative correction and completion.

The format for lightweight linguistic annotation has also been successfully tested on a sample of 46 texts of the Deutsches Textarchiv project (DTA 2018). The DTA consists of a large corpus of historical German texts, dating back to the 17th to 19th century. For normalization, lemmatization, and POS tagging the DTA applies the integrated system CAB (Jurish 2012) which provides various XML-based output formats, some of which already include the linguistic features discussed here in a similar way as the proposed format and may hence easily be converted to the format proposed here. Figure 11 and 12 illustrate this by example of the sentence : « Es iſt ja von Natur nicht gut/ daß der Menſch allein ſey [...] » (Abel 1699: 49).

```
<w t="Es" exlex="Es" errid="ec" msafe="1">
<moot word="Es" tag="PPER" lemma="es"/></w>
<w t="ift" exlex="ist" errid="25429"
msafe="1">
  <moot word="ist" tag="VAFIN"
lemma="sein"/></w>
...
<w t="Menſch" exlex="Mensch" errid="ec"
msafe="1">
  <moot word="Mensch" tag="NN"
lemma="Mensch"/>
</w>
...
<w t="ſey" exlex="sei" errid="57805"
msafe="0">
  <moot word="sei" tag="VAFIN"
lemma="sein"/>
</w>
```

Figure 11: XML (TokWrapFast)-Output by CAB[11]

```
<w norm="Es" pos="PPER"
 lemma="es">Es</w>
<w norm="ist" pos="VAFIN"
 lemma="sein">iſt</w>
<w norm="ja" pos="ADV" lemma="ja">
 ja</w>
<w norm="von" pos="APPR"
 lemma="von">von</w>
<w norm="Natur" pos="NN"
 lemma="Natur">Natur</w>
<w norm="nicht" pos="PTKNEG"
 lemma="nicht">nicht</w>
<w norm="gut" pos="ADJD"
 lemma="gut">gut</w>
<pc norm="/" pos="$("
 lemma="/">/</pc>
<w norm="daß" pos="KOUS"
 lemma="daß">daß</w>
<w norm="der" pos="ART"
 lemma="d">der</w>
<w norm="Mensch" pos="NN"
 lemma="Mensch">Menſch</w>
<w norm="allein" pos="ADV"
 lemma="allein">allein</w>
<w norm="sei" pos="VAFIN"
 lemma="sein">ſey</w>
<pc norm="/" pos="$("
 lemma="/">/</pc>
```

Figure 12: Output of CAB (see fig. 11) converted to TEI with *att.linguistic*

7. Limits of application

The mechanism introduced here is intentionally minimalistic: it serves to adorn tokenized text with the basic information labels essential for useful linguistic processing. It is not to be used for cases where multiple grammatical interpretations need to be listed and disambiguated, as in fig. 1 above. It is also far from optimal for handling multi-word units, especially if they involve discontinuity and/or changes in ordering, both of which can be found e.g. in separable German prefixes, illustrated below. The ideal usage scenario is where tokens (pieces of text) match word forms (as defined by ISO MAF[12]) one-to-one. This is not easy to achieve in natural languages, and therefore some repair strategies will usually be necessary.

The examples in fig. 13 and 14 illustrate two possible strategies of handling word forms which do not match tokens 1:1. In the German sentence *Ich stimme dir zu* (« I agree with you »), the base, infinitive form of the verb, is *zustimmen*, with the prefix attached to the verb. In some contexts, however, the prefix gets separated, yielding the correspondence between a single word form and two tokens. The typical way to handle this is by expressing the dependency between the tokens by modifying the repertoire of part-of-speech symbols – in this very case, the label « PTKVZ » of the STTS tagset[13] signals the prefix of a split form, so that the two parts can be reassembled at some higher level of representation (fig. 13). The approach taken in fig. 14 reassembles the morphological parts already at the level of tokens, and uses a convention whereby grammatical information describing the entire word form is represented on its first token.

```
<w pos="PPER" lemma="ich">Ich</w>
<w pos="VVFIN" lemma="stimmen">
stimme</w>
<w pos="PRF" lemma="du">dir</w>
<w pos="PTKVZ" lemma="zu">zu</w>
```

Figure 13 : Dependency between the prefix *zu-* and the verb stem *stimme* encoded indirectly, by means of a POS label (“ PTKVZ ”).

```
<w xml:id="t2" pos="VVFIN"
lemma="zustimmen" next="#t4">
 stimme</w>
<w pos="PRF" lemma="du">dir</w>
<w xml:id="t4" prev="#t2">zu</w>
```

Figure 14: A fragment of fig. 13 with the dependency captured at the level of markup

However, while the representation proposed here is able to handle mild deviations from the 1:1 correspondence between word forms and tokens, it is not sufficient for handling complex multi-word units or for syntactic description – these require more powerful descriptive mechanisms. Similarly, in systems which rely on the presence of `@norm` and use its content for further linguistic analysis, cases where a historical token corresponds to more than one normalized token may also turn out to be beyond the scope of lightweight descriptive mechanisms. Nevertheless, the fact that both archives mentioned in Section 6 have used the proposal described here successfully shows that many other projects can benefit from it.

[11] See http://www.deutschestextarchiv.de/demo/cab/ for the web service.

[12] While ISO specifications created outside the scope of the ISO-TEI liaison need to be purchased, ISO makes all definitions publicly viewable in the new ISO Online Browsing Platform : https://www.iso.org/obp/ui.

[13] http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html.

8. Summary and outlook

The goal of *att.linguistic* together with `@norm` is not to facilitate in-depth linguistic annotation, but rather to equip 'off-the-shelf' TEI in the very basic tools that linguists can use, and that non-linguists can safely add to their existing resources in order to enhance them. Where more elaborate analysis is needed, with explicit distinction between tokens and word forms and/or with hierarchical or dependency structures and the like, other TEI-based devices should be used.

Introduction of the `@norm` attribute will be the topic of a forthcoming LingSIG feature request directed at the TEI Technical Council.

9. Acknowledgements

We would like to thank the anonymous LREC reviewers for their comments on an earlier version of this paper. We are grateful to the TEI community at large, and the LingSIG community in particular, for numerous discussions at the stage of the preparation of this proposal and during the process of approval by the TEI Technical Council. Finally, many thanks are due to Peter Stadler, who acted as a liaison between the LingSIG and the TEI Council.

10. References

10.1 Research Articles

Bański, P. (2010). Why TEI standoff annotation doesn't quite work: and why you might want to use it nevertheless. In Proceedings of Balisage: The Markup Conference. Vol. 5 of Balisage Series on Markup Technologies. DOI:10.4242/BalisageVol5.Banski01.

Bański, P, Gaiffe, B., Lopez, P., Meoni, S., Romary, L., et al. (2016). Wake up, standOff!. TEI Conference 2016, Sep 2016, Vienna, Austria. ⟨hal-01374102⟩

Budin, G., Majewski, S., and Mörth, K. (2012). Creating Lexical Resources in TEI P5. *Journal of the Text Encoding Initiative* 3. DOI : 10.4000/jtei.522.

Haaf, S. (2016): Corpus Analysis based on Structural Phenomena in Texts: Exploiting TEI Encoding for Linguistic Research. In Nicoletta Calzolari (Conference Chair) et al.: Proceedings of the 10th LREC, pages 4365–4372, Portorož, Slovenia, May. ELRA. http://www.lrec-conf.org/proceedings/lrec2016/pdf/1154_Paper.pdf

Jurish, B. (2012). Finite-state Canonicalization Techniques for Historical German. PhD thesis, University of Potsdam. urn:nbn:de:kobv:517-opus-55789.

Ide, N. (1998). Encoding Linguistic Corpora. In Proceedings of the 6th Workshop on Very Large Corpora, pages 9–17, Montreal, Canada. http://www.cs.vassar.edu/~ide/papers/ces.wvlc.pdf .

Pose, J., Lopez, P., Romary, L. (2014). A Generic Formalism for Encoding Stand-off annotations in TEI. 2014. ⟨hal-01061548⟩

Schöch, Ch. (2016): Ein digitales Textformat für die Literaturwissenschaften. Die Richtlinien der Text Encoding INitiative und ihr Nutzen für Textedition und Textanalyse", *Romanische Studien* 4.

Stegmann, J., and Witt, A. (2009). TEI Feature Structures as a Representation Format for Multiple Annotation and Generic XML Documents. In Proceedings of Balisage: The Markup Conference. Vol. 3 of Balisage Series on Markup Technologies.
DOI: 10.4242/BalisageVol3.Stegmann01

10.2 (Language) Resource References

CLAWS TAGSET C8. Last changed - N.I.S. 14 Jan 2001. http://ucrel.lancs.ac.uk/claws8tags.pdf.

Deutsches Textarchiv (DTA). Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache. Berlin-Brandenburg Academy of Sciences and Humanities. Berlin 2007-2018. http://www.deutschestextarchiv.de/.

EEBO-TCP (2018): Early English Books Online. Text Creation Partnership. University of Michigan Library. www.textcreationpartnership.org/.

ISO 24610-1:2006. Language resource management, Feature structures, Part 1: Feature structure representation.

ISO 24611:2012. Language resource management, Morpho-syntactic annotation framework (MAF)

LingSIG "word attributes" project (2017): https://github.com/LingSIG/wordAttributes

NKJP: National Corpus of Polish – Narodowy Korpus Języka Polskiego (2008-2012). http://nkjp.pl/.

Text Encoding Initiative Consortium (eds.). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.3.0. Last updated 31st January 2018. http://www.tei-c.org/Vault/P5/3.3.0/doc/tei-p5-doc/en/html/.

TEI Elements : http://www.tei-c.org/release/doc/tei-p5-doc/en/html/REF-ELEMENTS.html

Universal Dependencies Tagset: POS tags. Version 2. http://universaldependencies.org/u/pos/ (2014)

WebLicht: https://weblicht.sfs.uni-tuebingen.de/ (2018)

10.3 Sources

Abel, H. K. Wohlerfahrner Leib-Medicus der Studenten. Leipzig, 1699. In Deutsches Textarchiv. URN: urn:nbn:de:kobv:b4-200905199878

Gottfried, J. L. Newe Welt Vnd Americanische Historien. Frankfurt (Main) 1631. In Deutsches Textarchiv. URN: urn:nbn:de:kobv:b4-200905199012

Marperger, P. J.: Der allzeit-fertige Handels-Correspondent. 4Th ed. Hamburg 1717. In Deutsches Textarchiv. URN: urn:nbn:de:kobv:b4-20887-0