Unary TEI Elements and the Token Based Corpus

Thomas Krause[1], Carolin Odebrecht[1], Amir Zeldes[1] and Florian Zipser[1,2]
1 Humboldt-Universität zu Berlin, 2 INRIA

The establishment of TEI as a standard for textual data generated outside of the narrow domain of corpus linguistics in history, literature, philosophy and more, has led to a fruitful integration of encoding vocabulary from different fields of interest, but at a necessary cost of a large stock of elements, heterogeneous interpretations of those elements, and limitations on the kinds of annotation combinations that a schema allows.

Meanwhile in corpus and computational linguistics circles, advances in the direction of generic, vocabulary agnostic graph based models of corpus representation have gained prominence (notable examples are PAULA, Dipper 2005 and GrAF, Ide & Suderman 2007, the latter recently canonized as part of the LAF standard in ISO 24615). Graph based annotation formats lend themselves to generic, reusable query architectures, but reduce all data to having the same ontological status. Specifically, corpora in corpus linguistics center on the concept of tokens, minimal technical units of linguistic analysis, which serve as textual anchors for higher annotations (either features of the tokens, like parts of speech, or higher structures, such as syntax trees). In this paper we would like to point out a specific subset of problems caused by this dissonance between the TEI model and the token-based corpus annotation graph. We will focus on the interpretation of unary XML elements, such as line or page breaks (e.g. `<lb/>`, `<pb/>`), and the representation of the underlying data structure in non-XML-based corpus query systems.

Unary elements present a particular challenge for a token based corpus, since they occur within the plain text of a TEI document, yet they cover no part of the text, as shown in Figure 1.

```
<p>
…
thu es in einen kolben zusam=men
<pb n="15" rend="Am beſten zu Diſtilliren."/>
/ vnddiſtillirt das waſſer
…
</p>
```

Figure 1. A unary XML element for page breaks in the RIDGES corpus. (http://korpling.german.hu-berlin.de/ridges/).

In many corpus formats and corresponding search engines, a corpus relies on an ordered sequence of tokens (e.g. in TigerXML and the corresponding TigerSearch for treebanks, Lezius 2002). Markup around tokens is commonplace in many corpora and is interpreted as a span annotation applying to the enclosed tokens (e.g. in the widely used IMS Corpus Work Bench, CWB, see Christ 1994). But the page break annotation in Figure 1 applies not *to* tokens but *between* tokens. Treating it as a token in itself is a possibility to eschew

the problem, but this creates a further difficulty, in that the last token on the first page and the first token on the second page no longer form a consecutive sequence if a query for two adjacent words is carried out.

In part, this problem is caused by the limitations of inline XML, which prohibits hierarchy conflicts. In this particular case, a paragraph `<p>` element encompasses text on both pages around the page break, meaning that a binary element encompassing each page would create a hierarchy conflict of the type `<p>…<page>…</p>…</page>`. We would like to suggest that this situation is sometimes an undesirable artifact of XML technology, while in other cases it concerns a meaningful distinction that must be handled in a special way. The `<pb/>` case above can be argued to belong to the first class. In effect, although one could theoretically mark up just the point at which a page break occurs, the intention of the annotator may be, more often than not, to say which words belong on which page. The attributes `n="15"` and the `rend` attribute which contains the running head for the page, both suggest that the element stands for the entire page. The answer to the question 'what page is word X on?' involves reference to the preceding `<pb/>` element, implying that the annotation somehow applies to each token on that page. In formats or systems that do not forbid hierarchy conflicts (e.g. CWB), the page annotation can simply be stretched from one `<pb/>` element to its predecessor, creating natural page spans. This has been done in the RIDGES corpus, which was first converted to PAULA XML (Dipper 2005) and then imported into ANNIS (Zeldes et al. 2009), as shown in Figure 2.

hi						hi
hi_font						antiqua
lang						lat
lb	lb					lb
lemma	Name	zum	gut	brauchen	.	unknown
norm	Namen	zum	besten	brauchen	.	TRACTATVS
norm_auto	namen	zum	besten	brauchen	.	TRACTATVS
p	p					
pb	pb					pb
pb_n	4					5
pb_rend	in header: Vorrede.					
pos	NN	APPRART	ADJA	VVINF	$.	NN
pos_cor	NN	APPRART	ADJA	VVINF	$.	NN

Figure 2. TEI pb and lb annotations interpreted as spans in the RIDGES corpus.

However in some other cases, unary elements stand between tokens in a much more integral sense. For example, figures within a text that contain no text themselves can simply stand between two paragraphs or tokens in the same paragraph, as shown in Figure 3.

```
/ gutes / kräfftiges Waſſer bekommen.<lb/></p>
…
<div type="utensils">
   <figure rend=" Drawing of two cut-out boards"/>
</div>
…
<head type="margin">Andere weg<lb/>waſſer zu di=ſtillirn. </head>
```

Figure 3. A unary figure element encompassing no textual tokens.

In this case, there is no sensible way in which the figure annotation can be 'stretched' to include tokens before or after the figure: the figure literally contains no tokens. However as already discussed above, in a token based corpus-linguistic corpus, adding a token to correspond to the position of the figure will interrupt the text flow, e.g. with respect to the adjacency of the surrounding tokens. This is a problem both for consecutive token search and for searches 'within n words', since the figure would occupy the position of a token, just like a word. Furthermore, query systems which display a search result context window of e.g. ±5 tokens would show one word too few for each figure found in the context (since figures take up space in tokens). In fact, many search systems which use a key-word in context (KWIC) type of view, such as CWB, WordSmith (Scott 2012) or EXMARaLDA's search tool EXAKT (Schmidt & Wörner 2009) would also display the figure token position or take the utterance part following the figure to be complete, disrupting the search result in a potentially undesirable way.

In order to solve these issues, we suggest a mechanism we refer to as 'alternative segmentations' of a corpus, which we have implemented in the latest version of ANNIS3. Segmentations are annotation layers that have a special status in being allowed to determine search precedence/adjacency, context size and the visualization basis for KWIC views of the data. Figure 4 shows only the textual tokens from Figure 3 in the KWIC view, with the position of the figure between the tokens shown in the grid below.

gutes / kräftiges Wasser bekommen . Andere Weg Wasser zu destillieren .

1603-AlchemistischePraktik (grid)

figure						figure								
figure@rend						Drawing of two cut-out boards								
head							head							
head@type							margin							
lb							lb		lb					
lemma	gut	/	kräftig	Wasser	bekommen	.	ander	Weg	Wasser	zu	destillieren	.		
norm	gutes	/	kräftiges	Wasser	bekommen	.	Andere	Weg	Wasser	zu	destillieren	.		
p	p													
pb	pb													
pos	ADJA	$(	ADJA	NN	VVINF	$.	ADJA	NN	NN	PTKZU	VVINF	$.		
tok	gutes	/	kräftiges	Waſſer	bekommen	.	Andere	weg	waſſer	zu	di=ſtillirn	.		

Figure 4. A unary figure element is interpreted as an annotated token with not text, and ignored by the segmentation layer norm, shown above the grid in a KWIC view.

Although the figure element is in fact between tokens in the grid view at the bottom of the figure, it is possible to pose queries to ANNIS that ignore the figure in the calculation of adjacency. For example, the layer called 'norm' has been designated as a segmentation layer, and the following query in the ANNIS Query Language (AQL) finds the position in Figure 4 despite the intervening 'figure' annotation by using the corresponding typed precedence operator '`.norm`':

```
"." & "Andere" & #1 .norm #2
```

Similar queries for n-m unit distance are also possible, e.g. using operators like `.norm,3,4` and any number of segmentation layers may be defined and used as units for context size determination and the KWIC view. In our view, defining and using segmentation layers is a promising way of mediating between the generic graph interpretation of corpus data and users' needs in treating some forms of annotation as reference units for purposes of search and visualization in corpus query systems.

References

Christ, O. 1994. A modular and flexible architecture for an integrated corpus query system. In: *Proceedings of Complex 94. 3rd Conference on Computational Lexicography and Text Research*. Budapest, Hungary, 23–32.

Dipper, S. 2005. XML-based stand-off representation and exploitation of multi-level linguistic annotation. In: *Proceedings of Berliner XML Tage 2005 (BXML 2005)*. Berlin, Germany, 39–50.

Ide, N./Suderman, K. 2007. GrAF: A graph-based format for linguistic annotations. In: *Proceedings of the Linguistic Annotation Workshop 2007*. Prague, Czech Republic, 1–8.

ISO 24612. Language Resource Management – Linguistic Annotation Framework (LAF).

Lezius, W. 2002. *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora*. Doctoral Thesis, Institut für maschinelle Sprachverarbeitung Stuttgart, Germany.

Schmidt, T./Wörner, K. 2009. EXMARaLDA – creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics* 19(4), 565–582.

Scott, M. 2012. *WordSmith Tools version 6*. Liverpool: Lexical Analysis Software.

Zeldes, A./Ritz, J./Lüdeling, A./Chiarcos, C. 2009. Annis: A search tool for multi-layer annotated corpora. In: *Proceedings of Corpus Linguistics 2009*. Liverpool, UK.