Corpus Analysis based on Structural Phenomena in Texts: Exploiting TEI Encoding for Linguistic Research

Susanne Haaf
Berlin-Brandenburg Academy of Sciences and Humanities (BBAW)
Jägerstr. 22/23, 10117 Berlin
E-mail: haaf@bbaw.de

Abstract

This paper poses the question, how linguistic corpus-based research may be enriched by the exploitation of conceptual text structures and layout as provided via TEI annotation. Examples for possible areas of research and usage scenarios are provided based on the German historical corpus of the Deutsches Textarchiv (DTA) project, which has been consistently tagged accordant to the TEI Guidelines, more specifically to the DTA ›Base Format‹ (DTABf). The paper shows that by including TEI-XML structuring in corpus-based analyses significances can be observed for different linguistic phenomena, as e.g. the development of conceptual text structures themselves, the syntactic embedding of terms in certain conceptual text structures, and phenomena of language change which become obvious via the layout of a text. The exemplary study carried out here shows some of the potential for the exploitation of TEI annotation for linguistic research, which might be kept in mind when making design decisions for new corpora.

Keywords: Corpus Linguistics, Corpus Analysis, Annotation, TEI Structures, Historical Linguistics

1 Introduction

The Text Encoding Initiative (TEI) started off with the goal to provide scholarly projects and libraries with an annotation format that would help "facilitate the creation, exchange, and integration of textual data in machine-readable form, for all kinds of texts, in every human language, from every historical or social context." (Burnard 2013: 5). Indeed, as of today, the TEI Guidelines (2016) have become a de facto standard for various ways of text processing, ranging from digital text editions, digital text archives and corpora to scholarly journals, ontologies, dictionaries, and even non-textual resources like music and graphics, from the representation of historical and modern manuscripts and prints to the annotation of born-digital texts (TEI projects 2016, TEI SIGs 2016). In addition, a large number of tools have already been provided to further explore TEI resources.

The benefit of TEI/XML encoding is of course straightforward: texts are provided in a free format which is based on the XML standard, well-documented and suitable as an archival format. While being a machine-readable format it allows for the rendering of texts for different contexts (e.g. WYSIWYG versions for reading, critical editions, etc., digital as well as printed versions). However, depending on the preferred annotation depth the application of TEI encoding to texts can become a quite elaborate and costly task. This may lead to the question of further innovative usage scenarios of TEI text encoding, apart from just rendering and archiving texts. In fact, TEI encoding has already provoked other innovative ways of re-use, coming from very different disciplines and areas of interest, such as e.g. the linking of TEI texts with regard to certain features (e.g. Dumont 2016), the transformation of TEI structures to RDF and their further semantic enrichment (e.g. Schrade 2016), the visualization of structures for different purposes (e.g. Pytlik Zillig 2015), and even the sonification of TEI structures (e.g. Emsley 2015).

This paper focuses on the area of corpus linguistics and raises the question of how linguistic research might benefit from the exploitation of TEI text structuring within (historical) corpora. Examples for possible areas of research and usage scenarios are provided based on the German historical corpus of the Deutsches Textarchiv (DTA) project, which has been consistently tagged accordant to the TEI Guidelines, more specifically to the DTA ›Base Format‹ (DTABf). In this context this paper also draws special attention to the DTABf as a TEI format originating from the field of corpus linguistics and suitable for this kind of research.

After a clarification on the terminology used in this paper, section 3 outlines the context of TEI usage within the field of corpus linguistics. Section 4 offers a short introduction on the data and annotation basis of the analyses described here. Subsequently, four different research scenarios are described and explained in section 5, showing how research on the development of text structures themselves as well as on lexical and syntactical phenomena in text structures can benefit from TEI encoding. Finally, section 6 concludes with a summarization of the results.

2 Terminology

In this paper three ways of TEI text annotation are distinguished:

Layout annotation: The encoding of structural text phenomena reflecting the layout and other formal aspects of a source text, such as bold/underlined/indented text, column print, figures and tables, forme work, etc.

Annotation of conceptual text structures: The encoding of text structures with regard to text type or conceptual purpose (from a content oriented perspective), such paragraphs, titles, poems, citations, etc.[1]

Linguistic annotation: The encoding of language features (from a structuralist point of view), such as tokens, lemmas, parts-of-speech, grammatical structures, types of speech etc.

[1] Note that the DTABf documentation follows this distinction between formal and conceptual annotation (DTABf-CTSA 2016, DTABf-LA 2016).

3 TEI-Annotation for Linguistic Research

Linguists are one important and large user community of the TEI, represented even by a separate Special Interest Group (TEI for Linguists 2016). As expected, one huge issue in this context are TEI solutions for linguistic annotation (e.g. Pose et al. 2014, Stührenberg 2012). Besides this narrower linguistic interest, corpus linguists share an interest in structural TEI annotation in terms of layout information or conceptual text structures: Large text corpora were created offering TEI encoding, such as DWDS, BNC, DTA, EEBO-TCP, DEREKO etc.[2] and proposals were made by linguistic scholars concerning the enrichment of the TEI guidelines for the annotation of certain text types (relevant for corpus linguistics) (e.g. Beißwenger et al. 2012, Schmidt 2011).

While it is rather straightforward to think of linguistic usage scenarios with regard to linguistic TEI encoding, the benefit of layout and conceptual text structure annotation for linguistic research may not always be evident. Thus, with the DTA corpus at hand, this paper provides examples of linguistic research questions where TEI encoding is a either involved or even a crucial factor.

4 Preliminaries: Data and Annotation Basis

4.1 Corpus: The Deutsches Textarchiv (DTA)

The current study is carried out on the corpora of the Deutsches Textarchiv project (DTA 2016).[3] These corpora contain historical German texts, originating mainly from the 17th to 19th century, which have been digitized to allow for linguistic studies on the development of the New High German language. The DTA data basis consists of a core corpus comprising a selection of texts from various text types and disciplines, digitized and annotated in the course of the DTA project, as well as a corpus of additional texts, curated from external projects, adapted to the DTA guidelines and integrated as extensions into the DTA (DTAE 2016, Thomas/Wiegand 2015).

The analyses of the following examples 5.2–5.4 have been based on the DTA corpora (core corpus and extensions) as they are publicly accessible via the corpus search option on the DTA website, comprising 2,818 works in total. It was reduced to the time span 1600 to 1900, because this has been the scope of the DTA project and there are only sporadic texts older or younger than that in the DTA corpus. Thus, the current study is based on a total number of 2,649 works (183,377,535 tokens) dating back to the 17th to 19th century. Example 5.5 was based on the download version of the DTA corpora.[4]

4.2 TEI Format: The DTA Base Format (DTABf)

The analyses described here exploit the TEI tagging according to the DTABf, as it was applied to the DTA corpora. The creation of the DTABf was necessary especially for interoperability reasons, since interoperability of corpus data is necessary to ensure the reliability of research results gained from a corpus.[5] Interoperability in this context does not only comprise homogeneity of quality and transcription practice but also homogeneity of linguistic, conceptual and layout tagging. Thus, the creation of an annotation format for the DTA which is fully compliant to the TEI Guidelines but reduces the TEI tagset and adds specifications on attribute and value level in order to prevent not only ambiguous tagging but also slips and lapses, was an important task. It has been a challenge, too, given that the corpus texts represent very different time periods and genres and are thus highly heterogeneous in layout and style (Geyken et al. 2011). In addition, the DTABf was not only applied to the DTA core corpus but to all DTAE texts which have been digitized in other project contexts based on individual project specific guidelines and formats. The result is a corpus of homogeneously annotated texts which can now be taken as a basis for corpus-based research.

The design of the DTABf was described in detail in Haaf et al. (2015); its status with regard to other established TEI formats was discussed in Geyken et al. (2013); specifics of the DTABf for the annotation of historical newspapers were shown in Haaf/Schulz (2014). Here only those features of the DTABf will be presented shortly which are of interest for the examples below.

The DTABf forms a TEI subset for the annotation of conceptual text structures as well as layout phenomena within historical prints. Conceptual text structures within the DTABf include e.g. chapters, titles, title pages, poems, notes, speech acts in drama, letters, citations, etc. In terms of layout annotation structures like e.g. figures, tables, lists, page breaks, typographical specifics (e.g. bold print, italics) etc. are included (cf. DTABf-LA&-CTSA 2016).

The DTABf tagset is grouped in different levels indicating the required annotation depth (DTABf-L 2016). Elements of the first 2 levels have been consistently applied to the DTA core corpus, level 3 elements are optional.

4.3 Linguistic Annotation & Corpus Query

The following examples use the results of different (automatic) analyses of linguistic phenomena within the DTA, such as tokenization, lemmatization, recognition of Parts-of-speech, and orthographical modernization (Jurish 2012, DTA-SW 2016).

Corpus queries based on those linguistic features as well as other structural phenomena were carried out mostly by usage of the DDC query language (Jurish et al. 2014), except in example 5.5 (see below for details).[6]

[2] For information on the respective encoding used cf. for BNC: Burnage/Dunlop 1993, for DTA: Haaf et al. 2014/15, for DWDS: DWDS 2016, for EEBO-TCP: EEBO-TCP Encoding 2016, for DEREKO: Lüngen/Sperberg-McQueen 2012.

[3] Deutsches Textarchiv (German Text Archive, DTA), funded by the German Research Foundation. Cf. Geyken et al. 2011, DTA 2016.

[4] Available under: http://www.deutschestextarchiv.de/download (version of 4th February 2016).

[5] Interoperability issues with the TEI tagset are discussed in Unsworth 2011. Ways of addressing this problem with the DTABf are explained in Haaf et al. 2015.

[6] Documentations of the DDC query syntax are available under:

5 Examples for Linguistic Research Based on Text Structures

5.1 Introduction

Information about the conceptual text structures can often be extracted from a printed source only by analyzing its layout. Bold centred text in large characters indicates a title which again marks the beginning of a chapter. Footnotes are printed in smaller letters below the main text block of a page, generally introduced by a horizontal rule, and contain remarks of the author or editor of a text. Text in smaller letters at the margin of a page is usually a marginal note, citations are surrounded by quotation marks, poem verses are printed below one another and are usually shorter than the page width would allow, etc. Because of these specifics in layout in many cases it is not complicated to apply tagging of conceptual text structures even without knowing the contents of a text. This way, projects creating large text corpora like the DTA are able to achieve the annotation of a wide range of conceptual text structures, as well.

Since the tagging of conceptual text structures is thus the result of an interpretation and somehow 'translation' of layout phenomena, it adds an actual additional value to the text compared to its printed source. Additionally, this information becomes machine readable within a text corpus, thus allowing for the consideration of conceptual text structures in quantitative corpus research. The following examples show scenarios where the consideration of conceptual text structures in corpus-based analyses may lead to insights on the development of language phenomena.

5.2 Example: Development of Titles

One area of interest in this context may be the conceptual text structures themselves, for instance the development of the style of titles: Titles in German texts of the 17th and early 18th century often appear as sentences or sentence-like structures (containing finite and finite auxiliary verbs; fig. 1). In comparison, titles in later texts are more similar to the ones used today, consisting of a couple of meaningful terms (nouns, adjectives) only (fig. 2).

This observation leads to the hypothesis that titles in German texts got shorter, more compact and less complex over time. A way to verify this hypothesis would be to search for titles (i.e. TEI element <head>) which contain finite or finite auxiliary verbs (indicating sentence-like structures). A corpus query is carried out on the DTA corpus using the DDC search syntax in order to analyze this phenomenon:

***Example Query*:**

[Q1] $p={VAFIN,VVFIN} with $con=/head/

(Query for finite and finite auxiliary verbs in TEI element //head)

The curve illustrating the chronological progress of the queried structure (fig. 3) clearly shows that the usage of verbs in titles decreases over time.

To ensure the result in fig. 3 the progress of verb usage within the whole corpus over time is calculated showing that it continuously remains on a high level (fig. 4).

Example Query:

[Q2] $p={VAFIN,VVFIN} with $con=/text/

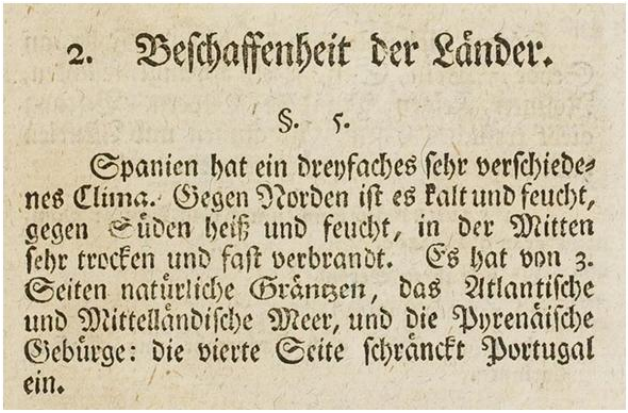
(Query for finite [auxiliary] verbs in <text> elements, i.e. the entire texts are considered containing all structures including titles)

The example illustrates that ignoring structural (layout) information for text annotation would mean to exclude important semantic and linguistic information about a text from the text transcription.


Wie die Obersten des Schiffs genent waren/vnd wo das schiff her war/vnd was sich noch begab ehe wir auß dem hauingen fuhren/vnd wie lang wir auff der reyse nach Franckreich waren. Caput liij.

Der Hauptman des schiffes war genant Wilhelm de Moner/vnd der Steurman Francoy de Schantz/ Das Schiff war genennet die Catharina von Wattaulla ꝛc. Sie rüsteten das schiff wider zů/nach Franckreich zusiegeln/so begab es sich eyns morgens/da wir noch in dem

Figure 1: Complex title (*Staden 1557: [119]*)


2. Beschaffenheit der Länder.

§. 5.

Spanien hat ein dreyfaches sehr verschiedenes Clima. Gegen Norden ist es kalt und feucht, gegen Süden heiß und feucht, in der Mitten sehr trocken und fast verbrandt. Es hat von 3. Seiten natürliche Gräntzen, das Atlantische und Mittelländische Meer, und die Pyrenäische Gebürge: die vierte Seite schränckt Portugal ein.

Figure 2: Compact title (*Achenwall 1749: 41*)

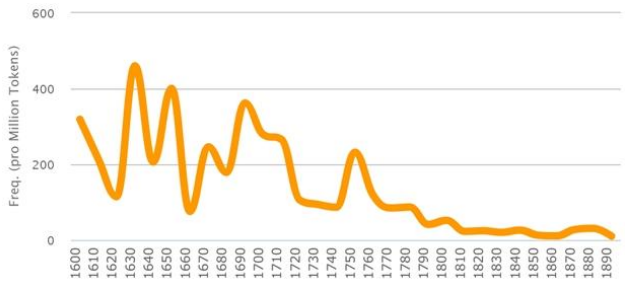


Figure 3: Finite (auxiliary) verbs in titles (result of [Q1])

http://www.deutschestextarchiv.de/doku/DDC-suche_hilfe;
http://odo.dwds.de/~moocow/software/ddc/querydoc.html.

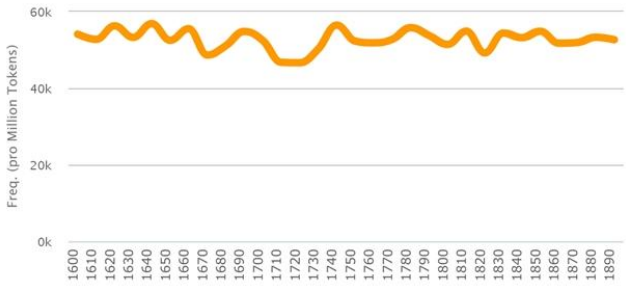


Figure 4: Finite (auxiliary) verbs in the entire DTA corpus

5.3 Example: Development of Marginal Notes

The results in Example 5.2 might lead to another research question: Historical texts often contain marginal notes attached to certain paragraphs. These marginal notes often summarize the contents of the paragraph or text passage they belong to (fig. 5). Thus, they seem to have a similar function like titles. If this is the case, the question arises if marginal notes develop over time in a similar way as titles with regard to complexity and length. Since margins have been annotated as such within the DTA corpus[7] it is now possible to perform a corpus query that helps examine margins with regard to this issue.

***Example Query*:**

[Q2] $p={VAFIN,VVFIN} with $con=/note_(left|right)/

(Query for finite and finite auxiliary verbs within right or left margins)

Interestingly, the results of this query show that margins do *not* behave similar to titles regarding complexity and length over time. The number of finite and finite auxiliary verbs in marginal notes is quite low compared to the number of marginal notes in general, indicating that there are far less marginal notes with complex syntax than compact ones during the whole time span the DTA corpus represents (Fig. 6&7). In addition, the curve for margins with complex syntax decreases similarly to the one for marginal notes altogether, showing the decreasing usage of marginal notes in printed texts over time in general.

A possible explanation for this observation is that in the case of marginal notes space on the printed page has always been an issue (as opposed to titles which may use the whole width of a page). In addition, the results show that margins do not as a general rule summarize the contents of the corresponding text passage or title it but frequently just contain bibliographical references for the statements within those respective text parts, e.g. biblical references within texts with theological contents (fig. 8).

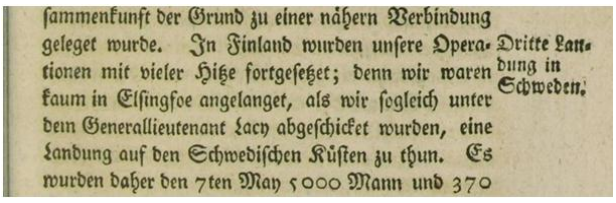


Figure 5: Marginal note (*Bruce 1784: 265*)

[7] Element <note> with @place="left|right". DTABf-CTSA 2016: 4.5.

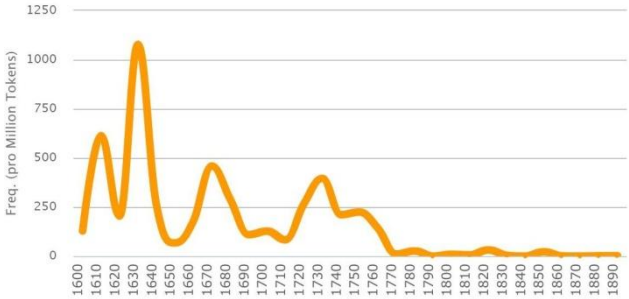


Figure 6: Finite (auxiliary) verbs in marginal notes

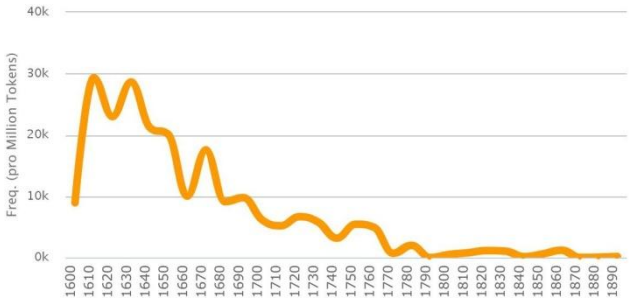


Figure 7: General development of marginal notes

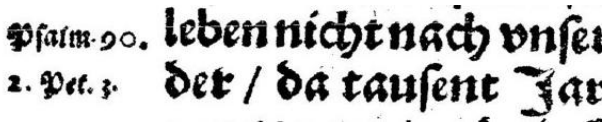


Figure 8: Marginal note (Aubelin 1602: 16)

5.4 Example: The Usage of *frohlocken* in Lyrics and Prose

Examples 5.2 and 5.3 showed how the annotation of conceptual text structures in corpora helps studying the (syntactical) development of those structures themselves over time. Another use case might be the comparative study of how single terms are used within different conceptual text structures.

An example is the German verb *frohlocken* (English: 'exult'). The Duden dictionary (1999) of the German language distinguishes three meanings of *frohlocken*:

1. *„Schadenfreude empfinden"*
   ('to sense malicious joy')
2. *„vor Freude jubeln; jauchzen"*
   ('exult', 'cheer of joy')
3. *„lobsingen (veraltet)"*
   ('praise/rejoice someone [archaic]')

Historical Dictionaries or those, concentrating on historical meanings, generally mention the second of those meanings: The Etymological dictionary of Pfeifer (1993/2016)[8] specifies *frohlocken* as *„lebhafte Äußerungen der Freude bekunden"* (i.e. 'to vividly express joy'). Adelung (1793-1801: 2,315)[9] refers to the meaning: *„einen hohen Grad der Freude so wohl durch die Stimme, als auch durch die Geberden an den Tag legen"* (i.e. ‚to show a high degree of joy using voice as well as gestures'). He translates *frohlocken* with Latin *"exultare"* (which

[8] http://dwds.de/?qu=frohlocken.

[9] http://www.woerterbuchnetz.de/Adelung?lemma=frohlocken.

brings to mind the English term 'exult'). The German Dictionary of Grimm (1854-1961: 4,227) [10] explains *frohlocken* with Latin *"jubilare"*.
What cannot be found in the dictionaries is if there are differences in the usage of the verb *frohlocken* with regard to the conceptual text structures it appears in. To find out more about this issue, *frohlocken* as it is used in poems can be compared to its usage in prose text. The DDC queries for this example are:

***Example Queries*:**

[Q3] frohlocken with $xpath=/\bp\b/
(Query for the term "frohlocken" within prose paragraphs)

[Q4] frohlocken with $xpath=/\bl\b/
(Query for the term "frohlocken" within poem verses)

The results show, that indeed, there is a significant difference of usage of *frohlocken* within prose versus lyrics. Among the 473 hits of *frohlocken* in prose this verb is mostly used without object or other complement (Usage 1: U1) or it is complemented by a prepositional phrase, commencing with the preposition *"über"* (i.e. 'about') (U2).

Manners of Use:

[U1] *frohlocken*
(without complement)[11]

[U2] *frohlocken über*
('exult about')[12]

The meaning of *frohlocken* in these cases is usually meaning 2 of the Duden dictionary, sometimes meaning 1 (cf. above).
Apart from these most common manners of use there are some other rare ones with only small amounts of hits.

Manners of Use:

[U3] *frohlocken mit den Händen*
('exult with hands'), 4 hits[13]

[U4] *jemandem entgegen/zu frohlocken*
('exult towards someone'), 2 hits[14]

[U5] *jemandem Beifall entgegen frohlocken*
('exult applause towards someone'), 1hit[15]

[U6] *in jem./etw. frohlocken*
('exult in so./sth.'), 3 hits[16]

In lyrics the absolute amount of hits for *frohlocken* is far smaller than in prose texts, only 64, which is explicable by the fact that the query performed included all prose texts within the DTA (fictional, functional and scientific) whereas lyrical texts are only a part of the sector of fictional literature.[17]
However, in lyrics there is one further, comparably frequent manner of use for *frohlocken*:

Manners of Use:

[U7] *jemandem frohlocken*
('praise someone'), 8 hits:

Result list:

[1] **Frolocket** eurem Held/ nach Streit hat er jetzt Ruh.
(Mühlpfort 1686: 48)

[2] **Frolockt** der Seeligen/ die nun von Angſt und Plage Entbunden
(Mühlpfort 1686: 89)

[3] **Frolockt** der Seeligen/ ſie iſt vergnůgt erhŏret/
(Mühlpfort 1686: 266)

[4] **Frohlockt** dir eine halbe Welt.
(Zachariae 1764: 254)

[5] Mir hůpfen die Berg', und die Hůgel! Mir **frohlocket** der Wald!
(Klopstock 1773: 147)

[6] Wiehert ihm Rosse! **frohlockt** ihm, Adler!
(Kosegarten 1798: 41)

[7] Ueber dem Haupt **frohlocken** ſie mir und ...
(Hölderlin 1826: 157)

[8] **Frohlockt** ihm, alle Frommen!
(Candidus 1854: 21)

Compared with the amount of hits in lyrics, U7 is quite frequent, comprising ~12.5% of the total amount of hits. It corresponds to the third meaning of *frohlocken* in the Duden dictionary (marked as archaic; cf. above). It can be found in texts of 6 different authors ranging from 1686 to 1854. This manner of use of *frohlocken* with dativ case complement can only be found in lyrics.

One might object that this analysis would not have required TEI tagging of the whole text but could have been performed by just including information on text categories given in the metadata records of the respective corpus texts. Such query would, however, have been far less accurate since on the one hand, in the case of lyrics it would have included non-lyrical text within volumes of poems, such as prose (prefaces, introductions, explanatory texts, notes), indices, title pages etc. On the other hand, poems contained by works of prose would have been evaluated as prose. By using the TEI annotation of the texts as a basis for the DDC query, it was possible to

[10] http://www.woerterbuchnetz.de/DWB?lemma=frohlocken, http://zwei.dwds.de/wb/dwb/frohlocken.

[11] *„Die Einŏde verwandelt ſich in ein Luſtrevier voll glŭcklicher Geſchŏpfe, die in Freyheit und Liebe frohlocken"* (Hirschfeld 1780: 53). *„Achilles frohlockte, als er ihn ſah."* (Schwab 1839: 265).

[12] *„der Himmel muß ſich auffthun und gleichſam frolocken ůber ſeiner Tauff"* (Dannhauer 1666: 274). With meaning 1 (sense malicious joy): *„Ja wann einer gefăhrlich kranck wird, ſo frolocken ſchon die Brůder ůber ſeinen verhoften Hinrichtungs-Tag,"* (Benner 1748: 247).

[13] *„Etliche frolockten mit Hănden/ ſchwungen die Můtzen ůber die Kŏpffe/ warffen ſie in die Hŏhe/ vnd fiengen Sie wieder/ ſchryen vnd jubilirten alſo vor vnd neben vns her"* (Olearius 1647: 291).

[14] *„Wir frohlockten ihm entgegen, aber ſeine Blicke waren nur auf die Ohnmăchtige gerichtet"* (Spieß 1796: 34/35).

[15] *„Der Teufel, der ihn nun zu ſeinem Zwecke vŏllig reif ſah, frohlockte ihm lauten Beyfall entgegen."* (Klinger 1791: 345).

[16] *„und ihre Seele frolockete in dem GOtt ihres Heyls"* (Meyer 1759: 151).

[17] The query is, however, valid since the study here focuses on manners of use rather than frequencies of use.

only include verse lines in Q4 and paragraphs in Q3 (excluding even poem titles, chapter titles, forme work, dates, indices etc.).

5.5 Example: Exploitation of Layout Annotation for the Study of Loanwords

The examples above showed usage scenarios based on the annotation of conceptual text structures in corpora. The deeper text structuring gets, though, the more difficult it gets to apply conceptual text structures at first sight. Examples are inline phenomena like e.g. proper nouns or foreign words. Thus, projects usually have to draw a line of annotation depth beyond which tagging, especially of conceptual structures, cannot be applied. For instance, for reasons of project resources and workflows the DTA project did not carry out the comprehensive annotation of proper nouns and foreign words in the DTA core corpus.[18] Nevertheless, in such cases where the application of conceptual structures was not possible, layout information found in the printed source was kept via the annotation. This way, since the layout of a printed source commonly reflects its conceptual text structures, it is still possible to find and explore conceptual text structures within the DTA core corpus via the specifics of layout. For instance, Antiqua print in gothic texts usually highlights foreign terms which have not been lexicalized, yet, or proper nouns (fig. 9). Thus, it is possible to query the corpus for Antiqua print in Fraktur texts in order to find proper nouns and foreign words, analyze them or even annotate them with the suitable conceptual tag.

In addition, layout tagging itself can lead to information about linguistic phenomena. Figure 10 shows how the process of lexicalization of a foreign term might become obvious by the way it was printed:[19] in the earlier texts (esp. 17th century) the German noun *Rarität* (English: 'rarity/curiosity') was printed in Antiqua typeface completely, a common method to highlight foreign words. During the 17th and 18th century, the Antiqua print of the foreign stem merged with Fraktur print of the suffix, a probable indicator of the lexicalization process of this term. In 19th century texts, usually the whole word is printed completely in Fraktur typeface indicating that the term had become adapted by the German language.[20]

... die sonnen/ vnd strecken sich nahe der	**Linea**	zu.
	Gelee	zu stürtzen,
.../ das er in seiner Gewalt hat/	**pupillariter**	substituiret/ ob dann wohl solche pupillaris **substitutio** ...
Ακρόνυχος	**enim**	**vespertinus est: qui oritur** ἄκρος νυκτὸς **primae** ...
auß seiner	**meditation**	vnnd betrachtung in der ersten **sermon,** den ...
einige Astronomen mehrere Stunden nach dem Austritt der	**Venus**	den Mond vor der Sonnenscheibe gesehen haben, ...
Tramseide, franz.	**Soyes**	**tremes,** nennet man diejenige Seide, die ...
Diesen Tag ist eine Griechische	**Polacro**	**,** ist ein Schiff von den Griechen also ...

Figure 9: Examples of Antiqua typeface in Fraktur texts

Figure 10: Typographical representation of term *'Rarität'*

6 Conclusion

This paper provided examples on how linguistic corpus-based research may be enriched by the exploitation of conceptual text structures and layout as provided via TEI annotation. It was possible to show that with this method significances can be observed in different linguistic areas, as e.g. the development of conceptual text structures themselves, the syntactic embedding of terms in certain conceptual text structures, and phenomena of language change which become obvious via the layout of a text.

The examples performed in this paper were based on the historical DTA corpus and the TEI annotation format DTABf. The DTABf, as a TEI format designed for the homogeneous annotation of historical text corpora, enables research as demonstrated here since it ensures the interoperability of annotated DTABf data and thus allows for reliable results of TEI-based corpus-research. This format has been created on the basis of the DTA corpus and was continuously further developed with the growth of the corpus. It has been successfully applied to more than 2,600 historical German works which shows its high coverage with regard to structural phenomena in historical texts. It is still maintained and further developed responding to user's requests and additional tagging necessities, usability issues or simply updates to the TEI Guidelines. In addition, the DTABf is the recommended format for the annotation of historical corpora in CLARIN (2016) and has also been recommended by the German Research Foundation in different contexts (DFG 2016).

Since TEI structuring is a time consuming task, providers

[18] Note that the DTABf provides tagging solutions for these phenomena, though. They are marked as level 3 elements, i.e. their usage is optional (DTABf-L 2016). Thus, it is possible to annotate proper nouns as well as foreign terms in a DTABf conformant way even though the DTA core corpus does not contain consistent annotation of these phenomena.

[19] Note that these analyses on the term *Rarität* were carried out on the dump of the DTA corpus as a whole (see above note 4), by usage regular expressions and XPath. Queries were:

a. [Rr]arit[eäåæ]t.* within XPath: //hi[@rendition='#aq'] ('*Rarität' within Antiqua typeface*),

b. <hi rendition="#aq">Rari.*?</hi>[\s\w] ('*Rarität' with Antiqua and Fraktur typeface alternating*),

c. [Rr]arit[eäåæ]t with XPath: //text()[not(parent::hi[@rendition='#aq'])] ('*Rarität' outside of Antiqua typeface*).

[20] For comparison: Foreign language terms in 19th century Fraktur texts are still printed in Antiqua typeface. Cf. e.g. Rabe (1803): 46.

of new corpora might question the necessity of TEI structuring for linguistic research. The exemplary study carried out here showed some of the potential for the exploitation of TEI annotation for linguistic research, which might be kept in mind when making design decisions for new corpora. Fortunately, since linguistic, TEI-based corpora already exist, further research leading to further examples in this area is possible.

7 Acknowledgements

I am grateful to my colleagues Matthias Boenig, Bryan Jurish, Christian Thomas, and Frank Wiegand for useful hints.

8 References

8.1 Bibliographical References

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L., Storrer, A. (2012). A TEI Schema for the Representation of Computer-mediated Communication. *jTEI* 3. http://jtei.revues.org/476

Burnage, G. and Dunlop, D. (1992). Encoding the British National Corpus. In Aarts, J.; de Haan, P.; Oostdijk, N. (Eds.), *English Language Corpora. Design, Analysis and Exploitation. Papers from the 13th International Conference on English Language Research on Computerized Corpora*. Nijmegen. http://www.natcorp.ox.ac.uk/archive/papers/Burnage93a.htm

Burnard, L. (2013). The Evolution of the Text Encoding Initiative. From Research Project to Research Infrastructure. *jTEI* 5. http://jtei.revues.org/811.

Deutsche Forschungsgemeinschaft (DFG). (2016). Förderkriterien für wissenschaftliche Editionen in der Literaturwissenschaft. http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/foerderkriterien_editionen_literaturwissenschaft.pdf

Dumont, S. (2016). correspSearch. Search diverse scholarly editions of letters. http://correspsearch.bbaw.de.

Emsley, I. (2015). "It will discourse most eloquent music". Sonifying variants of Hamlet. http://ora.ox.ac.uk/objects/uuid:1785e0ac-5cbb-4d35-8546-4495aa8baec8.

Geyken, A.; Haaf, S.; Jurish, B.; Schulz, M.; Steinmann, J.; Thomas, Ch. and Wiegand, F. (2011). Das Deutsche Textarchiv. Vom historischen Korpus zum aktiven Archiv. In Schomburg, I.; Leggewie, C.; Lobin, H. and Puschmann, C. (Eds.), *Beiträge der Tagung „Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland“, 2010*. hbz. pp. 157--161. http://www.hbz-nrw.de/dokumentencenter/veroeffentlichungen/Tagung_Digitale_Wissenschaft.pdf#page=159

Geyken, A.; Haaf, S. and Wiegand, F. (2013). The DTA ‘base format’. A TEI-Subset for the Compilation of Interoperable Corpora. In Jancsary, J. (Ed.), *Proceedings of the 11th Conference on Natural Language Processing (KONVENS)*. Vienna. http://www.oegai.at/konvens2012/proceedings/57_geyken12w/57_geyken12w.pdf

Haaf, S.; Geyken, A. and Wiegand, F. (2014/15). The DTA “Base Format”. A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources. *jTEI* 8. http://jtei.revues.org/1114.

Haaf, S. and Schulz, M. (2014). Historical Newspapers & Journals for the DTA. In *Proceedings of the LRT4HDA workshop, held at the 9th LREC Conference, 26–31 May 2014, Reykjavik*. pp. 50--54. http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-LRT4HDA%20Proceedings.pdf#page=57

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

Jurish, B.; Thomas, Ch. and Wiegand, F. (2014). Querying the Deutsches Textarchiv. In U. Kruschwitz, F. Hopfgartner, & C. Gurrin (Eds.), *Proceedings of the Workshop MindTheGap 2014: Beyond Single-Shot Text Queries. Bridging the Gap(s) between Research Communities* (co-located with iConference 2014, Berlin, 4. März, 2014), S. 25–30, 2014.

Lüngen, H. and Sperberg-McQueen, C. M. (2012). A TEI P5 Document Grammar for the IDS Text Model. *jTEI* 3. https://jtei.revues.org/508.

Pose, J.; Lopez, P. and Romary, L. (2014). A Generic Formalism for Encoding Stand-off annotations in TEI. <hal-01061548>. https://hal.inria.fr/hal-01061548.

Pytlik Zillig, B. L. (2015). Early Modern Drama. XSLT-based Visualizations of Tree and N-gram Sequences in XML Texts. http://tei2015.huma-num.fr/en/papers/

Schmidt, Th. (2011). A TEI-based approach to standardising spoken language transcription. In *jTEI* 1. http://jtei.revues.org/142.

Schrade, Th. (2016): XTriples. http://xtriples.spatialhumanities.de/index.html.

Stührenberg, M. (2012). The TEI and Current Standards for Structuring Linguistic Data. An Overview. In *jTEI* 3. http://jtei.revues.org/523.

Thomas, Ch.; Wiegand, F. (2015). Making great work even better. Appraisal and digital curation of widely dispersed electronic textual resources (c. 15th–19th centuries) in CLARIN-D. In Gippert, J. and Gehrke, R. (Eds.), *Historical Corpora. Challenges and Perspectives*. Tübingen. pp. 181--196.

Unsworth, J. (2011). Computational Work with Very Large Text Collections. Interoperability, Sustainability, and the TEI. In *jTEI* 1. https://jtei.revues.org/215.

9.2 Links & Websites

DTABf 2016: DTA Base Format. http://www.deutschestextarchiv.de/doku/basisformat

DTABf-CTSA 2016: DTABf. Structuring of Semantic (Meaningful) Phenomena. http://www.deutschestextarchiv.de/doku/basisformat_texterschliessung_inhaltlich

DTABf-LA 2016: DTABf. Structuring of Formal (Typographic) Phenomena. http://www.deutschestextarchiv.de/doku/basisformat_texterschliessung_formal
DTABf-L 2016: DTABf Levels. http://www.deutschestextarchiv.de/doku/basisformat_table?lang=en
DTAE: DTA Extensions. http://www.deutschestextarchiv.de/dtae
DTA-SW: Software in the DTA. http://www.deutschestextarchiv.de/doku/software
EEBO-TCP 2016: Early English Books Online Text Creation Partnership. blogs.bodleian.ox.ac.uk/eebotcp/
EEBO-TCP Encoding 2016. http://www.textcreationpartnership.org/why-sgmlxml-encoding/
TEI: Text Encoding Initiative. http://www.tei-c.org
TEI Guidelines: P5-Guidelines of the Text Encoding Initiative. http://www.tei-c.org/Guidelines/P5/
TEI projects 2016: Projects Using the TEI http://www.tei-c.org/Activities/Projects/
TEI SIGs: TEI Special Interest Groups (SIGs). http://www.tei-c.org/Activities/SIG/
TEI for Linguists 2016: TEI for Linguists SIG. http://www.tei-c.org/Activities/SIG/TEI_for_Linguists
DWDS: Das Digitale Wörterbuch der deutschen Sprache. http://www.dwds.de/

9.3 Language Resource References

DTA (2016). Deutsches Textarchiv. Grundlage für ein Referenzkorpus der Neuhochdeutschen Sprache. http://www.deutschestextarchiv.de.
Achenwall, G. (1749). Abriß der neuesten Staatswissenschaft der vornehmsten Europäischen Reiche und Republicken. Göttingen. In DTA. urn:nbn:de:kobv:b4-00905198268.
Aubelin, D. (1602). Leichpredigt. Tübingen. In DTA. urn:nbn:de:kobv:b4-200247-3.
Behrens, G. H. (1703). Hercynia Curiosa, oder Curiöser Hartz-Wald. Nordhausen. In DTA. urn:nbn:de:kobv:b4-200905199968.
Benner, J. H. (1748): Herrnhuterey in ihrer Schalkheit. Vol. 3. Gießen. In DTA. urn:nbn:de:kobv:b4-200905198391.
Brentano, C. (1838). Gockel, Hinkel und Gackeleia. Frankfurt. In DTA. urn:nbn:de:kobv:b4-20090519607.
Bruce, P. H. (1784). Des Herrn Peter Heinrich Bruce [...] Nachrichten von seinen Reisen in Deutschland, Rußland, die Tartarey, Türkey, Westindien u. s. f. Leipzig. In DTA. urn:nbn:de:kobv:b4-25080-4.
Candidus, K. (1844). Der deutsche Christus. Fünfzehn Canzonen. In DTA. urn:nbn:de:kobv:b4-20090519783.
Dannhauer, J. C. (1666). Catechismvs-Milch. Vol. 8. Straßburg. In DTA. urn:nbn:de:kobv:b4-200905198718.
Falke, J. v.(1858). Die deutsche Trachten- und Modenwelt. Ein Beitrag zur deutschen Culturgeschichte. Vol. 2. Leipzig. In DTA. urn:nbn:de:kobv:b4-200905191175.
Hirschfeld, Ch. C. L. (1780): Theorie der Gartenkunst. Vol. 2. Leipzig. In DTA. urn:nbn:de:kobv:b4-200905197747
Hölderlin, F. (1826). Gedichte. Stuttgart et al., In DTA. urn:nbn:de:kobv:b4-200905197104.
Klinger, F. M. (1791). Fausts Leben, Thaten und Höllenfahrt. St. Petersburg. In DTA. urn:nbn:de:kobv:b4-200905197124.
[Klopstock, F. G.] (1773). Der Messias. Vol. 4. Halle. In DTA. urn:nbn:de:kobv:b4-200905199504.
Kosegarten, L. G. (1798). Poesieen. Vol. 2. Leipzig. In DTA. urn:nbn:de:kobv:b4-200905199504.
Marperger, P. J. (1717). Der allzeit-fertige Handels-Correspondent. 4. Auflage. Hamburg. In DTA. urn:nbn:de:kobv:b4-20887-0.
Meyer, J. (1759). Die grossen und seligen Thaten der Gnade. Zürich. In DTA. urn:nbn:de:kobv:b4-25213-5.
Mühlpfort, H. (1686). Teutsche Gedichte. Vol. 1. Breslau et al. In DTA. urn:nbn:de:kobv:b4-20414-7.
Olearius, A. (1647). Offt begehrte Beschreibung Der Newen Orientalischen Rejse. Schleswig. In DTA. urn:nbn:de:kobv:b4-20442-1.
Raabe, H. A. (1803): Die Postgeheimnisse. Leipzig. In DTA. urn:nbn:de:kobv:b4-200905194731
Royer, J. (1648). Beschreibung des gantzen Fürstlichen Braunschweigischen Gartens zu Hessem. Halberstadt. In DTA. urn:nbn:de:kobv:b4-200905197931.
Schwab, G. (1839). Die schönsten Sagen des klassischen Alterthums. Vol. 2. Stuttgart. In DTA. urn:nbn:de:kobv:b4-200905195624.
Seckendorff, V. L. v. (1656). Teutscher Fürsten Stat. Frankfurt (Main). In DTA. urn:nbn:de:kobv:b4-20561-0
Spiess, Ch. H. (1796). Biographien der Wahnsinnigen. Vol. 3. Leipzig. In DTA. urn:nbn:de:kobv:b4-200905197410.
Staden, H. (1557). Warhaftige Historia und beschreibung eyner Landtschafft der Wilden/ Nacketen/ Grimmigen Menschfresser Leuthen [...]. [Marburg]. In DTA urn:nbn:de:kobv:b4-30037-6 .
Zachariae, J. F. W. [1764]: Poetische Schriften. Vol. 3. [Braunschweig]. In DTA.urn:nbn:de:kobv:b4-20676-2.

Adelung, J. Ch. (1793--1801). Grammatisch-kritisches Wörterbuch der Hochdeutschen Mundart mit beständiger Vergleichung der übrigen Mundarten, besonders aber der oberdeutschen. 2. Auflage. Leipzig. In Wörterbuchnetz. Trier Center for Digital Humanities. http://woerterbuchnetz.de/Adelung/.
Deutsches Wörterbuch von Jacob und Wilhelm Grimm (1854--1961). 16 Bände in 32 Teilbänden. Leipzig. In Wörterbuchnetz. Trier Center for Digital Humanities. http://woerterbuchnetz.de/DWB/.
Duden. Großes Wörterbuch der deutschen Sprache. (1999). 10 vols. http://www.dwds.de.
Pfeifer, W. (1993/2016). Etymologisches Wörterbuch des Deutschen. 2. forlaufend erweiterte und aktualisierte Auflage. http://www.dwds.de.