

**Univerzita Karlova v Praze, Filozofická Fakulta**

Ústav anglického jazyka a didaktiky

Studijní obor: Anglický jazyk

Diplomová práce



Vojtěch Kubánek

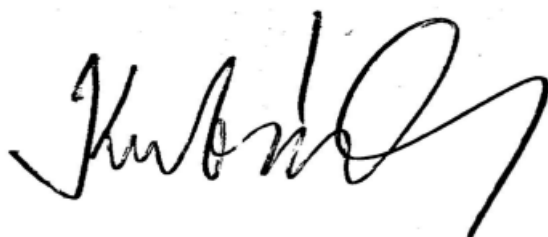
“Lexicographic Approaches to Old English”

“Lexikografické přístupy ke staré angličtině”

Vedoucí práce: Mgr. Ondřej Tichý, Ph.D.

2024

Prohlašuji, že jsem diplomovou práci vypracoval samostatně s využitím řádně ocitované literatury. Tato práce nebyla předložena jako splnění studijní povinnosti v rámci jiného studia nebo předložena k obhajobě v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

A handwritten signature in black ink, appearing to be 'Kubík' followed by a stylized flourish.

Rád bych poděkoval panu doktorovi Ondřeji Tichému za vedení této diplomové práce. Bez jeho cenných rad, věcných připomínek, a hlavně bez jeho předchozí práce na digitizaci tohoto slovníku by tato diplomová práce vůbec nemohla vzniknout.

## Abstract

The aim of this paper is to follow up on the digitization process of *An Anglo-Saxon Dictionary* originally written by Bosworth and Toller and now edited and digitized at the Faculty of Arts, Charles University under the lead of Ondřej Tichý. A short history of the dictionary will be followed by the description of the digitization process. The fine line will be explored between enhancing the original dictionary while staying true to it or recreating it into a new, different dictionary. One of the main topics will be the problem of preservation and standardization of digitized texts. A new standardized way of XML markup will be proposed that should facilitate the preservation of the digitized dictionary and make it inter-operable with other digitized dictionaries. Moreover, some changes to the current XML markup will be suggested that may further enhance the user-friendliness of the dictionary. The primary contribution of this paper is then a functional and valid TEI-Lex 0 markup for the structures found in the dictionary while supplementing all the necessary information for conversion between the current XML format and the standardized XML format. The secondary contribution is a unified markup in the current XML format for complex structures and suggested changes to the web application display that would further improve the user experience.

Keywords: An Anglo-Saxon Dictionary, digitization, Old English, lexicography, preservation, standardization, Bosworth, Toller

## Abstrakt

Cílem této diplomové práce je navázat na proces digitalizace Anglosaského slovníku, jehož původními autory byli Bosworth a Toller a který je nyní editován a digitalizován na Filozofické fakultě Univerzity Karlovy pod vedením Ondřeje Tichého. Po krátkém přehledu historie slovníku bude následovat popis procesu digitalizace. Bude zkoumána tenká hranice mezi vylepšením původního slovníku, který je však věrný své předloze nebo jeho přetvořením v nový, odlišný slovník. Jedním z hlavních témat bude problematika uchovávání a standardizace digitalizovaných textů. Bude navržen nový standardizovaný způsob značení XML, který by měl usnadnit uchování digitalizovaného slovníku a umožnit jeho vzájemnou kompatibilitu s jinými digitalizovanými slovníky. Kromě toho budou navrženy některé změny současného značení XML, které mohou dále zvýšit uživatelskou přívětivost slovníku. Hlavním přínosem této práce je pak funkční a validní značení TEI-Lex 0 pro struktury, které se ve slovníku nacházejí. Zároveň budou zmíněny všechny detaily pro konverzi mezi současným formátem XML a standardizovaným formátem XML. Sekundárním přínosem je jednotné značení v současném formátu XML pro složité struktury a návrh změn v zobrazení webové aplikace, které by dále zlepšily uživatelský komfort.

Klíčová slova: Anglosaský slovník, digitalizace, stará angličtina, lexikografie, uchovávání, standardizace, Bosworth, Toller

## **Abbreviations**

BT – *An Anglo-Saxon Dictionary* by Bosworth and Toller

DOE – *Dictionary of Old English*

DOEC – *Dictionary of Old English Corpus*

GLP – *Germanic Lexicon Project*

HTML – Hyper Text Markup Language

OE – Old English

XML – Extensible Markup Language

XSLT – Extensible Stylesheet Language Transformations

## Table of Contents

1. Introduction.....	9
2. Dictionaries of Old English.....	11
3. History of the Anglo-Saxon Dictionary .....	13
3.1. The Digitization process.....	14
4. Preservation of Digitized Texts.....	16
4.1. Text Encoding Initiative .....	18
4.2. TEI-Lex 0.....	19
5. Old English Dictionary Comparison .....	22
5.1 Number of Entries.....	22
5.2 Entry Content.....	23
6. User-friendliness and Fidelity to the Original.....	31
7. Formats of the Bosworth-Toller Dictionary.....	34
7.1 Format of the Printed Version .....	35
7.2 Current XML format.....	39
7.3 TEI Lex-0 XML Format .....	46
8. Non-prototypical and Complex Structures.....	58
8.1 Commentaries and Notes .....	59
8.2 Intra-example Glosses .....	63
8.3. Cognates, Reflexes, and Etymons .....	69
9. Further Changes .....	72
10. Summary.....	76
11. Shrnutí v Českém Jazyce .....	80
12. Bibliography.....	83

Figure 1 - A Dictionary Typology.....	11
Figure 2 - Technology Emulation Process.....	17
Figure 3 - Deprecated Collocate Structure.....	19
Figure 4 - TEI-Lex 0 Collocate Structure .....	20
Figure 5 - Entry structure of “dǣd-bētan” in <i>An Anglo-Saxon Dictionary Online</i> .....	24
Figure 6 - Entry Structure of “dǣd-bētan” in <i>Dictionary of Old English A to Le Online</i> .....	25
Figure 7 - Example Quotes of “dǣd-bētan” in <i>An Anglo-Saxon Dictionary Online</i> .....	27
Figure 8 - Example Quotes of “dǣd-bētan” in <i>Dictionary of Old English A to Le Online</i> .....	28
Figure 9 – Sense Numbering Error in “ge-sciftan” .....	31
Figure 10 – Comparison of Graphical Distinctions of “dǣd-bētan” in <i>An Anglo-Saxon Dictionary Online</i> and <i>An Anglo-Saxon Dictionary</i> .....	33
Figure 11 – Formats of the Bosworth-Toller Dictionary.....	34
Figure 12 – Structures Studied .....	36
Figure 13 – Related Entries Structure.....	37
Figure 14 – Etymology Structure .....	37
Figure 15 – Derivation Structure .....	38
Figure 16 – Editorial information in Supplement.....	38
Figure 17 – Current XML Structure of “stingan” .....	40
Figure 18 – Connectedness of Main Volume and Supplement Entries.....	42
Figure 19 – Graphical Distinctions of Parent Elements in the Web Application, .....	44
Figure 20 – Universal Current XML Structure.....	45
Figure 21 – TeiHeader Abridged.....	47
Figure 22 – Tei-Lex 0 XML structure of “stingan” .....	48
Figure 23 – Comparison of “pre-definition” Structure in Current XML Format and Tei-Lex 0 Format.....	49
Figure 24 – Comparison of Sense and Definition Structures .....	50
Figure 25 – Definitional Gloss in the Printed Version .....	51
Figure 26 – Usage Information in Definition in the Printed Version.....	51
Figure 27 – TEI-Lex 0 Markup of Usage Information in Definition.....	51
Figure 28 – Comparison of Example Structure.....	51
Figure 29 – Comparison of Etymology Structure .....	52
Figure 30 – Comparison of Related Entries Structure.....	53
Figure 31 – Comparison of Derivation Structure.....	53
Figure 32 – Conversion Table Between Current XML Format and TEI-Lex 0 Format.....	54
Figure 33 – Universal TEI-Lex 0 Structure.....	56
Figure 34 – Commentaries in the Printed Version .....	59
Figure 35 – Web Application Display of Commentaries.....	60
Figure 36 – Comparison of Commentary Structure .....	60
Figure 37 – Complex Commentary in the Printed Version.....	60
Figure 38 – Comparison of Simplified Markups for Complex Commentaries.....	61
Figure 39 – Comparison of Complex Markups for Complex Commentaries.....	61
Figure 40 – Further Complex Commentaries in the Printed Version .....	62
Figure 41 – Current XML Hybrid Markup of Complex Commentaries .....	63
Figure 42 – TEI-Lex 0 Hybrid Markup of Complex Commentaries .....	63
Figure 43 – Intra-example Glosses in the Printed Version .....	63
Figure 44 – Current Format Markup of Intra-example Glosses and Its Web Application Display .....	64
Figure 45 – Updated Current Format Markup of Intra-example Glosses .....	65

Figure 46 – TEI-Lex 0 Markup of Intra-example Glosses .....	65
Figure 47 – Intra-example Variant in the Printed Version, exemplified on “self” .....	66
Figure 48 – Current Format Markup of Intra-example Variants .....	66
Figure 49 – Updated Current Format Markup of Intra-example Variants .....	66
Figure 50 – TEI-Lex 0 Markup of Intra-example Variants .....	67
Figure 51 – Parenthesized Structures Typology .....	68
Figure 52 – Various Etymology Structures in the Printed Version .....	69
Figure 53 – Current Format Markup of Various Etymology Structures .....	70
Figure 54 – Updated Current Format Markup of Various Etymology Structures .....	70
Figure 55 – TEI-Lex 0 Markup of Various Etymology Structures .....	71
Figure 56 – Font-carrying Elements and Their Web Application Display .....	72
Figure 57 – Category-defining Elements and Their Web Application Display.....	73
Figure 58 – Sense-numbering Elements and Their Web Application Display .....	74
Figure 59 – Reference to a Specific Sense in the Printed Version.....	74
Figure 60 – Orthographic Variants of Cognates in the Printed Version .....	75



## 1. Introduction

*Would it not have been far better for Bosworth's memory to have let the good he did live after him, the evil lie interred with his bones, rather than to have thus raked up all the errors of the infant Anglo-Saxon scholarship of his time and republished them in this year of grace 1882, a confession of Englishmen's ignorance of the philology of their own tongue? (Platt, 1884, p. 237)*

It has been 140 years since the publication of Platt's critique of Bosworth and Toller's *An Anglo-Saxon Dictionary* (BT) and it did not age very well. More than a century after the publication of the dictionary, it is still regarded as the primary source of choice for researchers of Old English. This diploma thesis will describe this 140-year-long journey of BT, the various supplements added to the dictionary, the transition to the electronic age marking the digitization of the dictionary, and the possible future of the dictionary, proposing improvement both to the digitized version of the dictionary and to the preservation method.

The paper can be distinguished into three basic parts, the first being theoretical and ranging from Chapter 2 to Chapter 4.2. In this part, the typology of Old English dictionaries will be described and the type under which BT belongs will be given. After that, the publication history of BT and the digitization process that followed will be explored. The last chapters of this part will be devoted to the problem of the preservation of digitized texts and the ways in which standardization facilitates this task.

The second part ranging from Chapter 5 to 7.2 will deal with the description of the primary source, that is BT, and its various formats. First, BT will be compared to its supposed successor – *Dictionary of Old English* – with the emphasis put on the content comprehensiveness of each dictionary. The next section will disclose the main premise of the digitization process, that is to find the balance between fidelity to the original and user-friendliness of the modernized digitized version. The ways in which this balance is attained will also be described in this section. Lastly, an in-depth description of the two currently existing formats of BT will be given, starting with the structural analysis of the printed version followed by the digitized version marked up via custom elements in Extensible Markup Language.

The third part ranges from chapters 7.3 to 9 and functions as a synthesis of the two preceding parts. Based on the chapters on the preservation methods and the usefulness of standardization, a new format of BT is proposed. This format would be also marked up in XML, however this time the elements used would conform to the standard and thus facilitate the preservation and inter-operability of BT. Following the proposed process during which the standardized format

would be derived from the current format, some more complex structures of BT that, so far, lack unified markup and thus are unconvertable to the standard, will be explored. In these cases, a unified markup will be proposed that would both improve the current format and make conversion to the standard possible. The last chapter of this section will be devoted to the user-friendliness of the web application and how it can be improved through minor changes to either the XML document or the XSLT document that transforms the structure of XML to HTML.

## 2. Dictionaries of Old English

The aim of this chapter is to give a general overview of the typology of Old English (OE) dictionaries and dictionary-like documents and to determine under which type Bosworth-Toller's *An Anglo-Saxon dictionary* (1898) – the primary source and focus of this paper – belongs. The first step is to find out where are OE dictionaries categorized in perspective to other types of dictionaries, for this, the basic typology as proposed by Zgusta (1971) and schematized by Swanepoel (2003) presented in the figure below, will be suitable:

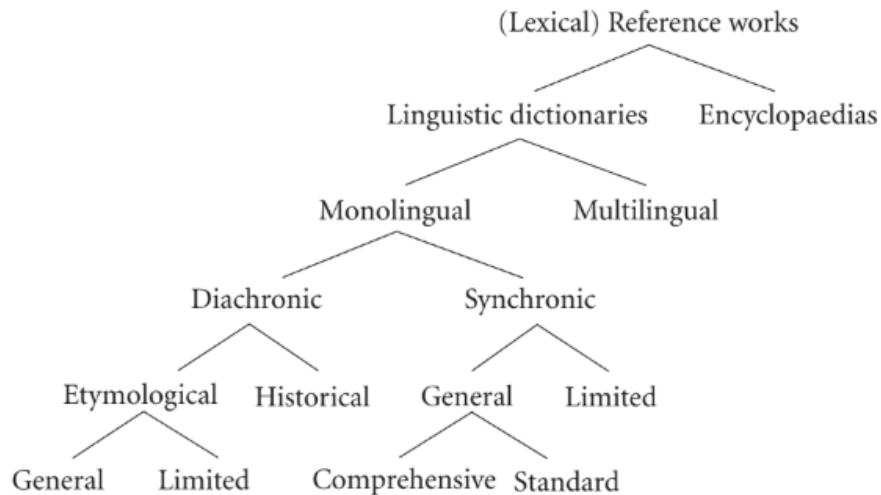


Figure 1 - A Dictionary Typology, based on Zgusta (1971, p. 198-221), schematized by Swanepoel (2003, p.46)

As per the figure, it is clear that OE dictionaries belong to the diachronic-historical and diachronic-etymological. Progressing further into the typology of diachronic dictionaries, several other distinctions can be made. The further categorization is based on the dictionary types found in *A Practical Guide to Lexicography* (Sterkenburg ed., 2003) where the other types discussed and relevant in terms of OE lexicography are dictionaries of authors and texts, restricted dictionaries, pedagogical dictionaries, standard dictionaries and lastly comprehensive dictionaries.

Starting with dictionaries of authors and texts, this term is used interchangeably in *A Practical Guide to Lexicography* (Ibid.) with the more common term “glossaries”. This type of OE dictionary is restricted to a single text or author and its function is to give additional information on the specific terms and vocabulary of the primary text(s). One such glossary can be found in Baker's *Introduction to Old English* (2012) on pages 283-387, where the vocabulary described is taken from texts in the “Anthology” on pages 181-275 and other chapters of the book. Next up is the wide category of restricted dictionaries, whose scope of restriction can range anywhere from grammatical limitations (e.g. dictionaries of nouns) to functional limitations (e.g. dictionaries of abbreviations). In terms of OE lexicography, one of the representatives of the

restricted dictionary category<sup>1</sup> is the *Dictionary of Old English Plant Names* (Bierbaumer et al eds., 2007-2009). Moving on to pedagogical dictionaries that are, as the name suggests, devised with the teaching function as the basis of the dictionary. For OE lexicography one of the most influential of such students' dictionaries was Sweet's *The Student's Dictionary of Anglo-Saxon* (1896) – this dictionary is often compared to its contemporary and even more influential counterpart *A Concise Anglo-Saxon Dictionary* (Hall, 1894) which, however, could be categorized under both pedagogical and standard dictionary types. In terms of OE lexicography, it may be better to use only the term standard dictionaries as the boundaries between pedagogical and standard dictionaries are blurry. Although even better term may be concise dictionaries as per the title of Hall's oeuvre, as it makes the categorical distinction to comprehensive dictionaries clearer. Comprehensive dictionaries typically offer the same number of entries as concise dictionaries; however, the depth as to which the entry headwords are described is more profound, and hence the length of the entries is larger. The most known representative of the comprehensive OE dictionary category is the primary research object of this paper – *An Anglo-Saxon Dictionary* (Bosworth and Toller, 1898). Another, as impactful, yet currently not finalised, comprehensive OE dictionary, which is generally seen as the successor of BT is the *Dictionary of Old English* (DOE) (Cameron et al eds. 1986-). For a comparison and further description of the two comprehensive OE dictionaries mentioned, please see Chapter 5. Now, that it is clear to what category BT belongs in comparison to other OE dictionaries, let us turn to the publication history of BT and the almost 200-year period that led to its current digitized format.

---

<sup>1</sup> More precisely, the category of content restricted dictionary.

### 3. History of the Anglo-Saxon Dictionary

The publishment history of BT is a convoluted one. The first abridged edition written solely by Bosworth was published in 1838 under the name *A Dictionary of the Anglo-Saxon Language*. This version of the dictionary was considerably less comprehensive than its successors, the reviews for this edition were unfavourable, some reviewers going as far as calling it a “botch” (Baker, 2003, p. 95). The second edition, written largely by Toller based on Bosworth’s notes and manuscripts was given the slightly changed title *An Anglo-Saxon Dictionary* and was published in its entirety in 1898. However, before the publication of the entire dictionary, it was published as four distinct volumes – parts I and II, published in 1882, ranging from A to H were still for the most part written by Bosworth (A-F and most of G solely by Bosworth (Ibid., p. 96)) and H written by Toller as was the rest of the dictionary in parts III and IV. The reviews of the first two parts were, again, rather critical: “[T]he continuation of the work by Toller appears to be almost as bad as the commencement of it by Bosworth” (Platt, 1884, p. 237). The full edition<sup>2</sup> was not treated as harshly, however, some criticism remained mainly regarding the (lack of) lemmatization, obsolete orthography, or the way of referencing primary sources (Garnett, 1898). Toller was aware of the problems pointed out by the critics during the publishment of the Main Volume and therefore already in 1898 promised to make amends in the supplement to the dictionary, although, pointing out that the creation of a faultless dictionary of OE is a task near impossible (Toller, 1898, preface). This work titled *Supplement to An Anglo-Saxon Dictionary* was first published in 1921 and its main purpose seems to have been the elimination of the lemmatization problem as many headwords – orthographic variants of the same lemma – from the Main Volume were grouped under a single headword and rest removed, yet the two other aforementioned problems have remained even in the supplement as the contemporary critic Schlutter mentions in his review (1919)<sup>3</sup>.

Based on the contemporary reception of each version of the BT, it may seem that after the publication of the supplement, the work on the dictionary would end. An overall better, more comprehensive, easier to navigate, and more structurally consistent dictionary was bound to take its place but as it turned out, Toller was right in his assessment of the difficulty of such a task and since then, there has not been a completed dictionary of OE so comprehensive or impactful as BT. As time progressed the reception of BT was getting more favourable, Ellis (1993, p. 4-5) claims that: “the Bosworth-Toller dictionary is far superior to Bosworth’s earlier work, and together with Toller’s 1921 *Supplement*, this work remains the most comprehensive

---

<sup>2</sup> Henceforward referred to as the “Main Volume”.

<sup>3</sup> The review was written before the official publication of the entire supplement.

Old English dictionary currently available.” During the century-long period, BT received further amendments out of which the most notable is Campbell’s “Enlarged Addenda and Corrigenda” (1972) utilizing 50 years’ worth of progression in Anglo-Saxon studies: “part is from newly published [OE] sources [...] and part from recent re-interpretation of long-known texts” (Page, 1975). However, it is not a part of the digitized dictionary as it has not become a public domain yet.

The digitization of BT<sup>4</sup> began in 2001 as a part of the *Germanic Lexicon Project* whose leader and founder was Sean Crist. The project was hosted on the GLP website until 2006, the website<sup>5</sup> is still functional and the pre-2006 digitized version where the many people responsible for the digitization are all duly accredited is still to be found there. In 2006 the project was taken up by the current leader of the digitization process Ondřej Tichý and it was transferred to servers hosted by Charles University and finally, in 2010 it was transferred to the website at <<https://bosworthtoller.com/>> where it has been hosted ever since. What exactly is the digitization process and how has it been done regarding BT will be the topic of the following chapter.

### 3.1. The Digitization process

First, the term digitization needs to be described as it is sometimes used interchangeably with other terms such as digitalization or digital transformation. Digitization in this paper describes: “the creation of digital artifacts through technical processes of conversion, representation, and enhancement” (Gradillas and Thomas, 2023), where the digital artifact is the electronic version of the dictionary which is created through the conversion, representation, and enhancement of the original BT. The digitization process consists of several steps that may be different based on the digitized medium. This study will be preoccupied with the digitization of texts generally, and with the specific digitization of BT. All the information regarding BT digitization is based largely on Tichý and Roček’s paper “Bosworth-Toller’s Anglo-Saxon *Dictionary* Online” (unpublished).

The first step in the digitization of any text is to convert the printed letters (the analogue) to machine-readable data (the digital). This can be done in a number of ways, the simplest but the least economical would be to manually retype the letters into a text document – a practice useful for smaller projects but unthinkable for texts of more than a thousand pages such as BT. The,

---

<sup>4</sup> That is the main volume and supplement

<sup>5</sup> Link to the website: <[http://www.germanic-lexicon-project.org/texts/oe\\_bosworthtoller\\_about.html](http://www.germanic-lexicon-project.org/texts/oe_bosworthtoller_about.html)>

by far, most widely used method, which has also been used for BT, is the scanning of pages<sup>6</sup> and further conversion of these images into the digital via Optical Character Recognition (OCR). During OCR, all the printed characters are standardized into a singular form<sup>7</sup> which amongst all the benefits it brings may also create some discrepancies between the original text and the digitized text as some characters may be recognized and in turn transcribed erroneously. As for BT, this part of the digitization process was still done during the leadership of Sean Crist in 2004.

Some of digitization processes may end at this point, but in terms of usefulness, they would bring a very limited number of new features. Therefore, in order to give the public a useful tool, several other steps were taken, this time, already with Ondřej Tichý as the leader. The second step, taking place in 2010, was the transformation of the data in line with the Unicode standard, which was especially helpful for special characters such as runic symbols used in the original. This in turn facilitated the transfer of the data to a document where a further description of the text in Extensible Markup Language (XML) could be given. In 2013, the first structures were marked up using automated scripts based on graphical signals as distinctions of various structures<sup>8</sup>. However, as the critics of BT pointed out in the preceding chapter, the structure of BT is inconsistent, and thus automated tagging could not have been used as the sole markup method. In 2016, a custom schema made to accommodate all the inconsistencies of BT was developed and the manual tagging of structures began and has continued until the present. So far, the last step in the digitization process was the creation of a new website by Martin Roček in 2021, the technicalities regarding the processing of the data and the specifics of the website's functionalities are beyond the scope of this paper, for further information of this topic, please see (Tichý and Roček, sec. 5). What, however, is not beyond the scope, is the way of preserving such digitized texts and the importance of inter-operability with other similar texts, both attained through the conformation to standardized formats.

---

<sup>6</sup> This again can be done manually page-by-page or in an automated, yet more financially demanding, way.

<sup>7</sup> In printed medium, all graphemes are innately (ever-so slightly) different whilst in digital form, they are the same.

<sup>8</sup> For a close description of the structures and their graphical distinctions, see 7.2.

#### 4. Preservation of Digitized Texts

Preservation, when referring to tangible objects, is the act of stopping the deterioration of that given object and maintaining or improving its current condition so that the object may benefit future generations. This idea of benefit – of being useful – is what differentiates preservation from conservation. But what does that imply to the digital medium? There is not a physical copy of BT deteriorating as each user metaphorically turns the pages in the web application. When we talk about digital preservation what we have in mind is the: “preservation of access [where] preservation is the action and access is the thing” (Conway, 2000, p. 16). Therefore, it is not an act of preserving the data but rather access to the data, where the meaning of access is two-fold.

The first meaning of the word entails access to a specific place where the data are safely stored – in the past, they were stored on CDs or DVDs which then had to be preserved in the “tangible-object meaning” of the term. Nowadays, the most common way of storage is through digital repositories that are mostly made of national and institutional digital archives. One such archive is the LINDAT/CLARIAH-CZ long-term data preservation repository where the data for BT are stored.<sup>9</sup> The second meaning involves the storage of accessible data, i.e. data that can be accessed by the future generations. To give an example, an XML document stored at a digital archive with no metadata information and vague elements such as <x> or <mmm> whose real functions are known only to the author cannot be preserved as with the death of the author, the text becomes unreadable and practically useless. To avoid this, Lee et al. (2002) suggest four main techniques of digital preservation: “technology preservation, technology emulation, information migration, and encapsulation”.

Technology preservation entails the preservation of the data, and all other software and hardware needed to access this data - a technique nowadays perceived as deprecated due to space limitation among other problems. Technology emulation is similar to the preceding technique with some extra steps. The period software and hardware need not be preserved if the data are given an emulation description - a sort of metadata used to convert the obsolete data format to the new format. How this process functions can be seen in the figure by Rothenberg (2000):

---

<sup>9</sup> The link to the storage can be found here: <<http://hdl.handle.net/11234/1-3532>>



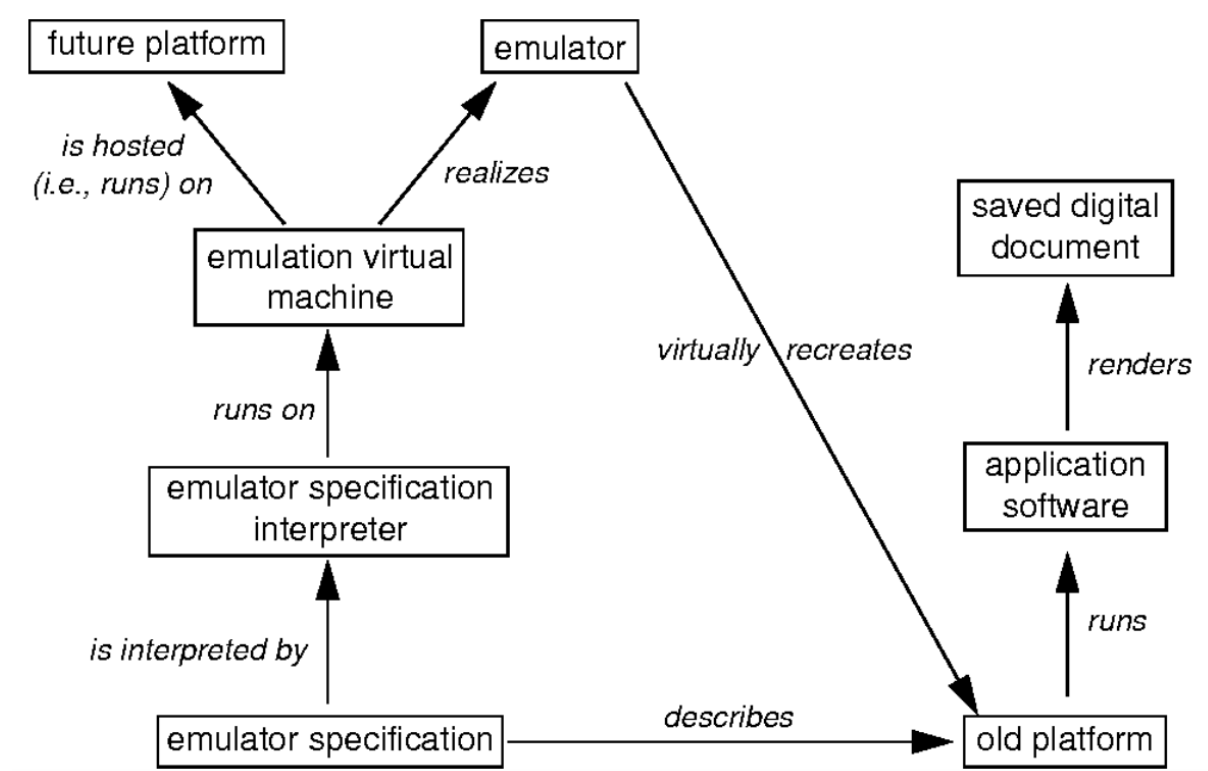


Figure 2 - Technology Emulation Process (Rothenberg, 2000)

The third technique of information migration entails the idea of recurring conversion of data from the obsolete format to the new format. Unlike the previous technique, the aim is not to emulate the outdated format through new software but rather to periodically convert the data as new technologies arise. The last technique is that of encapsulation which binds both the data and information on how to read the data together creating a self-sufficient capsule. This method is mainly used for stable objects whose structures do not change.

In terms of BT, the preservation technique is a mixture of migration and encapsulation as the XML data is bundled together with its schema, XSLT document, and metadata but as it is a living object being actively accessed and changed it bears also some of the notions of migration. However, in terms of XML, there is another very important concept which is that of a standardized markup: “The most successful preservation strategies will contain elements of migration based on standardization.” (Lee et al., 2002, sec. 4). That is because the information on how to access the data will be the same for all the documents and so will be the format to which the data will be converted once the current format becomes obsolete. The most widely used XML standard, at least for digital humanities, is the Text Encoding Initiative (TEI), the subject of the following chapter.

#### 4.1. Text Encoding Initiative

Text Encoding Initiative is a project developed for the long-term preservation and interoperability of XML documents following this standard. Its roots can be traced to the year 1987 and since then it has become the most recognized standard for digital humanities. The most important part of TEI is its schema against which all TEI-conformant documents have to be validated. The easiest way to find out whether a document is TEI-conformant is to validate it against the basic schema by including its namespace in the first element of the XML document like so: “<TEI xmlns="http://www.tei-c.org/ns/1.0">”. However, as extensive as the current TEI markup is, it cannot accommodate to the needs of all particular texts, therefore, the user is encouraged to transform the schema to suit their project’s needs. For this, a specialized tool was developed – the Roma editor. In Roma, the user restricts the elements needed for his project either through choosing specific modules or specific elements, e.g. when marking up a dictionary, the user can (and should) omit elements devised specifically for drama, this can either be done by omitting the whole “drama module” or by omitting singular elements from the module such as <camera> used to tag camera angles during a recorded play. Once having only the required elements, if an element is missing for a structure found in the project’s text, the user can generate new elements, although, in order to claim his schema to be TEI-conformant, a precise description (metadata) of the function of the element must be given. When all this is done, the user’s schema will be given a unique namespace, they will gain access to the full documentation of the schema and will be able to download the schema in one of the various formats (DTD, RNG, Schematron, etc...) and from then on, they will be able to mark up their text in a TEI-conformant way.

The second condition for a document to be TEI-conformant is to follow the TEI guidelines. Having a valid document against a TEI schema is not enough to call your document TEI-conformant as, for example, you may have chosen to mark up the lemmas of your dictionary by <camera> as it is shorter than <form type="lemma">. Document with such markup would be valid against the TEI-schema but it would not be TEI-conformant as it does not follow the guidelines. In the guidelines, every element is assigned a specific function and a specific place in the XML hierarchy<sup>10</sup> to ensure that all the projects are inter-operable (i.e. to prevent a lemma being in one project tagged as <camera>, in second by <item expand="lemma"> and in third by, for instance, <p>) and thus worth of preservation. However, at least for dictionaries, the guidelines are still not rigid enough and whilst in most cases the markup of two independent

---

<sup>10</sup> For XML „parent-child“ hierarchy see chapter 7.2.

dictionary projects would be similar, it very often would not be the same, thus undermining the basic idea of a standard. Fortunately, this problem has been solved by the TEI-Lex 0 community who devised a more constraining “sub-standard” of TEI for the specific needs of dictionaries.

#### 4.2. TEI-Lex 0

The TEI-Lex 0 project is a much younger sibling to TEI, with its beginnings being traced to 2016. The main aim is to overcome what Tasovac (2017), one of the leaders of the project, terms different “TEI flavours” which make the TEI’s promised inter-operability unfeasible. To facilitate interchange and inter-operability, the TEI-Lex 0 team created a new set of more constraining guidelines derived from the TEI guidelines. The flavours of TEI-valid markup are distinguished into deprecated structures and TEI-Lex 0 conformant structures; therefore, all TEI-Lex 0-conformant documents are also TEI-conformant documents but not vice versa. To give an example from the guidelines (sec. 3.3.3):

```
<entry>
  <form>
    <orth>médire</orth>
  </form>
  <gramGrp>
    <colloc>de</colloc>
  </gramGrp>
</entry>
```

Figure 3 - Deprecated Collocate Structure (Tasovac et al., 2018, sec. 3.3.3)

This kind of markup is TEI-conformant but not TEI-Lex 0-conformant. It is one of the many deprecated structures as the markup can be done in a number of valid ways<sup>11</sup>. TEI-Lex 0 comes with a singular standardized way of marking up structures found in dictionaries such as the example above of a grammatical morpheme in collocation with the lemma:

---

<sup>11</sup> E.g. instead of <gramGrp>, one could have <list type=“grammar“>, instead of <colloc> there could be <item type=“collocation“>, etc...

```

<entry xml:lang="fr" xml:id="DDL.F.médire">
  <form type="lemma">
    <orth>médire</orth>
  </form>
  <gramGrp>
    <gram type="collocate">de</gram>
  </gramGrp>
</entry>

```

Figure 4 - TEI-Lex 0 Collocate Structure (*Ibid.*)

The idea is simple, if all projects follow this markup instead of the various TEI-conformant markups, inter-operability between dictionaries will finally be attainable.

Looking at the figure above, another important idea of TEI-Lex 0 is presented, that is the idea of most precise markup as Tasovac (2017, 8:00-8:12) puts it: “[T]he more effort you put into encoding a legacy dictionary, the more useful it will be as a resource for semantic, linguistic, historical, cultural research.” In comparison to TEI, TEI-Lex 0 makes use of a more restricted set of elements, which, on the other hand, are further distinguished by mandatory attributes based on the type of structure, e.g. instead of marking up every form of a word as <form>, TEI Lex-0 distinguishes these forms as <form type=“lemma”>, <form type=“variant”>, or <form type= “inflected”>. Moreover, it is also deprecated to leave out defined structures from the markup, i.e. if TEI-Lex 0 offers markup for a structure found in the dictionary you are working with, this structure has to be tagged, e.g. if a dictionary makes use of special symbols known as metamarks<sup>12</sup> they have to be part of the markup as in this way the highest degree of inter-operability with other dictionaries using metamarks is ensured. The last advantage of TEI-Lex 0 over TEI is the fact that everything is stored on a single platform as customizations are no longer possible. All TEI-Lex 0 documents are validated against a single schema and are given the same documentation, out of which both are stored alongside the guidelines on the TEI-Lex 0 website.

All in all, TEI-Lex 0 has been found to be the best choice when it comes to the preservation and inter-operability of digitized dictionaries. But what does that mean in respect to BT? As mentioned earlier, currently the digitized version of BT uses custom elements and is validated against a schema tailored for the specific needs of BT which is however: “relatively easily transformable into TEI” (Tichý and Roček, unpublished). From this springs one of the main

---

<sup>12</sup> Metamarks are graphic symbols with a specific function in a particular text, in BT, the most common metamark is “:---” with the function of distinguishing between the definition and example categories.

foci of this paper, to take this transformation one step further and devise a plan for conversion of the current custom-tagged data into the standardized markup of TEI-Lex 0 – a task further described in chapter 7.3. But for now, another question has to be answered: Is it really worth it to preserve BT? Have not there been dictionaries in the last century that surpassed BT making them a better alternative for digitization and preservation? These questions will be answered in the following chapter comparing BT to another as impactful and newer OE dictionary.

## 5. Old English Dictionary Comparison

The comparison of BT will be held against the *Dictionary of Old English* (DOE) edited by Cameron, Crandell Amos, DiPaolo Healey, et al., it is the newest addition to the portfolio of OE dictionaries, with the beginnings of the project tracing to 1969 and its first part, consisting of entries under the letter D, being issued in 1986 (Jenkins, 1991). As of today<sup>13</sup>, more than half of the dictionary has been published ranging from letter A to Le and the work on the rest is currently under way. The aim of DOE is to be the most comprehensive source of OE lexicography which is attained through connection to the *Dictionary of Old English Corpus* (DOEC) which consists of all currently discovered OE sources which means that DOEC comprises the full data record of OE lexis and DOE gives the most comprehensive lexicographic description based on that data – a method far more advanced than was technologically possible during the creation of BT. This is why DOE is generally seen as the successor of BT once the work has been finalized. The following comparison thus serves not only as a tribute to the progression of OE lexicography but also as a practical assessment of whether BT will be of any use once DOE is published in its entirety.

### 5.1 Number of Entries

Starting off with some raw data, looking, for the sake of conciseness, only at entries listed under the letter D, at first glance it may seem that BT is the more comprehensive with the total number of entries being 1768 (955 main volume, 813 supplement), in comparison to DOE's 733.<sup>14</sup> There are several reasons for this discrepancy, starting with the most obvious, the supplement of BT mainly consists of entries already mentioned in the main volume with some information being added, deleted, or substituted. Out of the first 200 supplementary entries, 80 of them<sup>15</sup> were new unique entries and 120 were editing entries. The exact number of unique entries under D is not important but taking the numbers out of the first 200 entries studied<sup>16</sup>, the number would be somewhere between 1300-1400 which is still a considerably larger number than of the DOE.

The second reason for the discrepancy is the way the headwords are treated. As mentioned in Chapter 3, BT has, since its conception, been criticized for the lack of lemmatization, i.e. several orthographic forms of the same lemma are treated as separate entries. For instance, forms

---

<sup>13</sup> The 4th of August 2024.

<sup>14</sup> For a fair comparison, all entries starting with “ge-” have been omitted as in BT, they are listed under G. For DOE this is 191.

<sup>15</sup> Out of which 17 were suffixes.

<sup>16</sup> Cca 40% of the supplementary entries treated as unique entries.

meaning “day” such as “dæg”, “doeg”, “daga”, and many others are listed as separate entries in BT whilst DOE groups them all under the lemma “dæg”. In order to give a definitive answer as to which of the dictionaries has the most comprehensive overview of OE lexis, all entries starting with “da-” and “dæ-” from each dictionary were studied and compared. If there was an entry headword in any of the dictionaries that was missing from the other, it was checked whether it is listed as an orthographic form, and if it was not, only then it was categorized as dictionary-specific. In DOE, there have been four dictionary-specific headwords: “dalisc”, “dalland”, “dægbōt”, and “dægwilla” whilst in BT such headwords amounted to 16 entries. Ten of these entries were suffixes<sup>17</sup> and six were proper nouns: “Dægsan stán”, “Dærenta-múpa”, “Dalamensan”, “Datia”, “Daðan”, and “Dauid”. It is a common practice for dictionaries to either exclude proper names (as done in DOE) or list them in the appendix, however, BT may have been ahead of its time as: “[the] current practice is to include all headwords in one single list” (Atkins and Rundell, 2008, p. 179). As for affixes, the common practice is to exclude them from the headword list with the exception of productive affixes (Ibid., p. 166-167). Therefore, this practice is carried out better by DOE as it lists the productive “-dæda” (doer) and excludes the other 10 bound suffixes listed by BT.

All in all, the comparison of the number of entries uncovered some important facts about both dictionaries. While, at first glance, BT seems to offer the more comprehensive overview of OE lexis by a large margin, this notion is quickly dissipated as we find out about the imperfect lemmatization. Regarding common nouns, DOE is the better source as its connection to a complete corpus of OE sources is unmatched by BT, however in singular cases even BT may include common nouns missing in DOE<sup>18</sup>. On the other hand, when it comes to certain lexicographic decisions as, for example, the inclusion of proper nouns, BT does a better job even in terms of the current best practice and gives the fuller picture of OE lexis. Yet, the number of entries is only one of the indicators of the comprehensiveness of a dictionary with the second, as important if not more, being the actual content of the entries.

## 5.2 Entry Content

This chapter will compare the entry content of a single medium-length entry “dæd-bētan” upon which the general differences between BT and DOE content will be shown. The general

---

<sup>17</sup> All the suffix entries of BT under “da-” and “dæ-” with the exception of “-dæda” which is listed also in the DOE

<sup>18</sup> One of such examples is the entry “delfin” in BT meaning “dolphin”. The most common word for dolphin in OE was “mereswin”, however “delfin” seems to have entered English already during the OE period as a Latin loanword which is reflected in BT but not in DOE.

differences shown will be only amplified in larger and more convoluted entries, however, due to the impracticality of presenting longer figures only the relatively short “dǣd-bētan” will go through the close scrutiny, however if the entry does not show a certain difference sufficiently, longer entries will be referred to. The web application view of “dǣd-bētan” in BT can be seen here with the DOE counterpart right below:<sup>19</sup>:

# dǣd-bētan

Verb [ weak ]

 Dictionary links

[OED](#) [NED](#) [MED](#) [DOE](#) [DOEC](#) [PIE](#)

 Grammar

dǣd-bētan, part. -ende ; p.-bētte ; pp. -bēted

- I. To make amends, give satisfaction, to be penitent, to repent; *maleficium compensare, malum bono pensāre, pœnitere*

[Show examples](#)

 Linked entries

v. bētan.

Figure 5 - Entry structure of “dǣd-bētan” (An Anglo-Saxon Dictionary Online, 2014, <https://bosworthtoller.com/7312>)

<sup>19</sup> The example section can be found in figures below.



## dǣd-bētan

Vb., wk. 1, intrans.

Att. sp.: dǣdbetan || dǣdbetst || dǣdbete, dǣtbete | dedbete (HomU 23 MS E) || dǣdbeten | dedbetan || dǣdbetende || dǣdbætende (Æ) || dǣdbetendum, dǣþbetendum (xii).

Wk.: dǣdbetenden, dǣdbetendan, dǣdbettendan (BenR) | dędbetendan (BenR MS F)

ca. 30 occ.

1. to repent

1.a. ecclesiastical: to do penance, atone, literally ‘to make good or to remedy a deed’

1.b. in general sense: to regret, repent

2. *dǣdbetende*, present participle

2.a. as adjective

2.a.i. repenting, atoning

2.a.ii. *dǣdbetende sealmas*, glossing *psalmi paenitentiales* ‘the penitential psalms’

2.b. as substantive: a penitent, one who repents or atones

Lat. equiv. in MS: paenitere, satisfacere; paenitentialis = *dǣdbetende*

See also: *dǣd*, *bētan*; *dǣdbēta*, *dǣdbētere*; cf. *dǣdbōt*; *dǣde betan* s.v. *dǣd* sense 3.a.i., *dǣdum betan* s.v. *dǣd* sense 3.b.

Figure 6 - Entry Structure of “*dǣd-bētan*” (Dictionary of Old English A to Le online, 2024)

Starting at the top, both dictionaries give the information on the parts of speech and the categorization into weak/strong classes<sup>20</sup> (if the headword is a verb), DOE also includes the classification into classes as per Campbell’s *Old English Grammar* (1959) and the (in)transitivity of the verb. The second category for both dictionaries is the word-form category, where each dictionary chooses a slightly different method. BT gives some of the attested orthographic variants<sup>21</sup> and theoretical inflected forms based on the grammatical environment, for instance, as per Figure 5, a theoretical past participle form of the lemma would be “*dǣd-bēted*” (however such form has not been attested yet as can be seen in DOE). DOE, on the other hand, works only with the attested spellings from the DOEC, with grammatical information being given very sparsely. In general, this category will be more comprehensive in DOE as it

<sup>20</sup> The printed version of BT does not entail this information, for more information see chapter 7.2.

<sup>21</sup> No orthographic variants are part of this entry in BT but if they were, they would be listed in this category.

lists every single form of the lemma, however, the trade-off is, at least at present<sup>22</sup>, that the grammatical information given is insufficient. The next category, contained solely by DOE, is the number of occurrences. This makes sense as DOE, in comparison to BT, claims that it contains all the occurrences, this information thus gives the user a general idea of the frequency of use of the word.

The biggest discrepancy between the two dictionaries is to be seen in the sense and example category. Throughout BT, the sense category is much simpler than the DOE's, although, it has to be noted that Toller's part of the main volume and supplement does contain more convoluted sense structures and thus mitigates the difference between BT and DOE a little bit (cf. entry "habban" in both dictionaries – BT 55 senses vs DOE 151 senses). Therefore, it can be easily declared that in these terms DOE is more granular and hence more comprehensive (whether comprehensible I will let the reader decide). The second category mentioned is the example category, in which DOE, again, outshines BT. Due to DOE's access to all written OE lexis, the amount of examples given in an entry can be equal to the number of attested occurrences described above, yet, this would prove impractical in certain situations (see "habban" 12700 occurrences), as for the entry at hand DOE lists 19 examples against BT's three. Although, that does not always mean that the example section of BT is just a less comprehensive version of the DOE example section as can be seen in the figures below:

---

<sup>22</sup> The "entry format document" at the DOE website lists grammatical information as pertaining to the attested spellings category, yet as of now, very limited grammatical information is given even for lengthy entries such as "beran" or "gyfan", however for some entries entries such as "habban" the information is already there.

- I. To make amends, give satisfaction, to be penitent, to repent; *maleflicium compensare, malum bono pensāre, pœnitere*

Hide examples

His sǣwle wūnda **dǣdbētende** gelācnian  
to heal the wounds of his soul by making amends, Homl. Th. i. 124, 14.

**Dǣdbēte**

shall make amends, L. C. S. 41; Th. i. 400, 16: L. Eth. ix. 26; Th. i. 346, 6.

Ðæt he sealde sōðe gebȳsnunge eallum **dǣdbētendum**, ðe to Drihtene gecyrrap  
that he should give a true example to all, who shall turn to the Lord by doing amend deeds,  
Ælfc. T. 38, 4.

Figure 7 - Example Quotes of “dǣd-bētan” (An Anglo-Saxon Dictionary Online, 2014, <https://bosworthtoller.com/7312.>)

1. to repent

1.a. ecclesiastical: to do penance, atone, literally ‘to make good or to remedy a deed’

**HomU 27** 214: god wyle swaþeah gemiltsian æghwylcum synfullum menn, þe his synna her andet his scrifte and **dǣdbetan** wyle and æfre geswican þæs unrihtes, þe he ær worhte and dyde.

**BenR** 44.7.30: be þam amansumedan, hu hi **dǣdbeten** (*de his qui excommunicantur, quomodo satisfaciant*; BenR 44.70.2 *dǣdbetan*, BenRW 44.93.16 *hu hi sculon don hyra dadboten*).

**BenR** 11.36.6: gif hit gelimpe, **dǣdbete** [later glossator adds *satisfaciat*] se Gode on his gebedhuse, þe hit þurh his gymeleste gelamp (*satisfaciat deo in oratorio*; BenRW 11.47.22 *dǣdbete*, BenRGl 11.42.6 *gebete*).

**BenR** 26.50.10: gif hwylc broðor butan his abbodes hæse gedyrstlæcð, þæt he on ænige wisan ænige geþeodrædene nime wið þone amansumedan ... sy he gelicum gelimpe amansumad, and on gelicre wrace **dǣdbete** (*similem sortiatur excommunicationis uindictam*; F *dætbete*, OT *dædbote*; BenRW 26.67.8 *dedbore*, taking vb. as noun).

**BenR** 44.70.17: forðon on þa wisan mid hreowsunge **dǣdbete**, oð hit þam abbode fulbet þince and hine geswican hate (& *sic satisfaciat*; BenRW 44.95.2 *dedbore*, BenRGl 44.79.5 *fulgebete & hiht*).

**BenRW** 44.95.4: ða þe for litlum gylte fram gemenum <geswustra> gereorde ascyrede beod, þa on cyrcan þæslicre **dedbetan** & þæt fuldon on þare abbodesse hese, oððæt heo hi bletsige & secge, hit is genoh (*satisfaciant*; BenR 48.70.21 *dǣdbetan*, BenRGl 48.79.8 *hit gebeta*).

**Lit 5.11.7** 36: geswicenesse ic behate and æfter þinre tæcinge **dǣdbetan** wille mid eadmodlicre onhnigenesse.

**LawVIIIAttr** 26: gif mæssepreost manslaga wurðe ... þonne þolige he ægðores ge hades ge eardes & wræcnige swa wide swa papa him scrife & **dǣdbete** georne (LawIICn 41 *dædbete*, A *dætbete*).

1.b. in general sense: to regret, repent

**LibSc** 65.18: *sine consilio nihil facias et post factum non peniteberis* butan geþeahhte naht þu do & æfter dæde þu na **dædbetst**.

2. *dædbetende*, present participle

2.a. as adjective

2.a.i. repenting, atoning

**ÆCHom I, 4** 211.159: se apostol þa bebead þam twam gebroðrum þæt hi þrittig daga behreowsunge **dædbætende** Gode geoffrodon (T *dædbote*; cf. **Vir.Iohan.** 2.58.34 *ut per triginta dies pœnitentiam offerrent*).

**ÆHom 6** 250: <ac> ðæra synna <forgyfenyss stent on þam halgan> gaste, and he deð forgyfenysse ðam <**dædbetendum** mannum> (B *dæþbetendum*).

**BenR** 45.71.5: gif hwylc broðor wægð and misfehð on boduncge sealma ... butan he þærrihte beforan eallum hine **dædbetende** geeaðmede, he stiðran and teartran steore underfo (*nisi satisfactioe ibi coram omnibus humiliatus fuerit*; **BenRW** 45.95.12 *dædbetende*, **BenRGl** 45.79.12 *þurh fulre dædbote*).

2.a.ii. *dædbetende sealmas*, glossing *psalmi paenitentiales* ‘the penitential psalms’

**RegCGI** 65.1567: *sacerdos ... cum reliquis illius ministris misse ... eant ad uisitandum infirmum, canentes psalmos poenitentiales* se sacerð ... mid oþrum þære þenum mæssan ... gan to geneosigenne þære untruman singende sealmas **dædbetende**.

2.b. as substantive: a penitent, one who repents or atones

**ÆCHom II, 4** 38.270: ure drihten Iesus Christus. se ðe is soð sacerð gelæt þa **dædbetendan** æfter soðre dædbote to ðære uplican Hierusalem.

**ÆCHom II, 5** 49.222: uton besceawian ða micclan Godes arfæstnysse. hu he urum gyltum miltsað. and ðærtocan þæt heofenlice rice behæt. soðlice **dædbetendum** æfter gyltum (cf. **GREG.MAG. Hom.evang.** 19, 152.227 *caeleste regnum paenitentibus etiam post culpas promittat*).

**ÆLet 4 (SigewardZ)** 1149: he sealde soðe gebysnunge eallum **dædbetendum**, þe to drihtene gecyrrað, þæt hig magon arisan, gif hig rædfæste beoð, fram heora sawle deaþe & fram heora <synna> bendum, & heora scippend gladian mid soðre dædbote.

**Alc** 383: ne seceð God na swa swyðe þa lenge þære tide, swa he sceaweð þære lufe smyltnysse þæs **dædbetenden**.

**BenR** 28.7.8: hu abbod careful beon sceal ymbe þa **dædbetendan** (*qualiter debeat abbas sollicitus esse circa excommunicatos*; **BenR [F]** 27.50.17 *dædbetendan*, **BenRW** 27.67.11 *amansumadan*).

**ChrodR 1** 33.1: fram Eastron oð Pentecosten tuwa on dæg etan preostas, and etan flæsc be leafe, butan þa **dædbetendan**, buton Wodnesdæge and Frigedæge (*nisi penitentes*).

Figure 8 - Example Quotes of (Dictionary of Old English A to Le online. 2024.)

The first example in BT cannot be found in DOE which, considering the relatively short nature of the entry, can serve as a precedent for all further entries. The precedent being that the DOE example section cannot serve as a simple substitute for BT’s section, but rather, in order to get

the most comprehensive overview, the example section of both dictionaries should be studied together. The figures above also unveil another important topic, that of translations. Whereas BT supplies the majority of its OE examples with a translation in either PDE or Latin, DOE does so in a much smaller measure only for texts that come from bilingual manuscripts – in that case, the DOE example contains both the Latin original and its OE translation or vice versa, but for sources written solely in OE, no translation is given to the reader which may prove to be a hindrance for the usefulness of DOE in comparison to BT.

For the last two sections, let us return to the bottom of Figures 5 and 6, where we can find the Latin translation equivalent section in DOE with no categorical counterpart in BT, however, the information contained in this DOE's section has already been contained in the sense definition by BT and in terms of the actual content, both dictionaries seem to be similarly comprehensive. The last section serves as a list of references to other similar entries. As can be seen from the figure, this section is, again, generally more comprehensive in DOE.

To summarize, in the grand scheme of things, DOE is definitely the more comprehensive of the two dictionaries offering fuller information in the grammatical section, orthographic variants and inflected forms section, sense definition and examples section, and lastly even in the referential section. Notwithstanding all these advantages however, it cannot be taken as a simple substitute for BT as it lacks (as of now) information on the grammatical environment of the inflected forms, some of the OE examples listed by BT, and the majority of BT's translations. Moreover, it lacks some of the useful proper noun entries in BT and is generally not as accessible to the general public as BT is. If these reasons are still not enough to convince the reader of the desirability of BT's preservation, one could point out the other indisputable motives, BT is a historical artefact showing the state of OE lexicography in the 19<sup>th</sup> and 20<sup>th</sup> centuries, whilst it may not be the most lexicographically up-to-date dictionary, its "redundant" structures may serve as encyclopaedic information<sup>23</sup> or details of the word-formation<sup>24</sup> processes in OE.

All in all, it has been evaluated that for the most comprehensive description of OE, BT and DOE should be studied together and therefore the preservation and accessibility of OE is a task worth pursuing. But before we get into the standardized markup of TEI-Lex 0 that would be of great service to this task, first we need to define in which way we want to preserve the dictionary,

---

<sup>23</sup> BT, mainly in parts written by Bosworth, sometimes gives lengthy encyclopaedic comments (see. "dæg" or "Bēowulf").

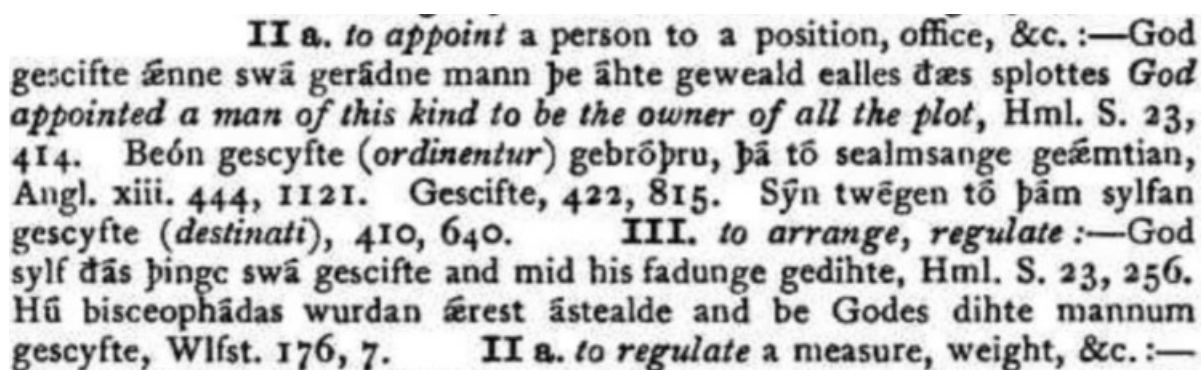
<sup>24</sup> See the suffix commentary above.

either as a historical artefact with all its disadvantages or as a modern improved version of itself  
– as a useful tool.

## 6. User-friendliness and Fidelity to the Original

As mentioned in Chapter 4, enhancement of the digitized object is an integral part of the preservation process. However, there is a fine line between enhancing and recreating which will be different based on the project. In this chapter, the aim will be to list some of the features added and modernized in terms of the web application BT as opposed to the original BT. For features that may benefit (either from a lexicographical or user-centred point of view) from modernization but were left true-to-original, the reasoning behind this decision will be given. It has to be kept in mind that absolute fidelity to the original is a task practically unattainable as only the fact that the analogue medium was changed into digital is already quite a drastic change. The task is then to balance the number of changes made for the sake of user-friendliness with enough features being left unchanged so that it is clear that the digital object is indeed the digitized version of BT rather than a new digital dictionary based on BT.

The most important part of the dictionary, that is the text, is left unchanged unless there is an apparent error. Such errors sometimes arise in the sense category, more precisely, in the numbering of the senses as can be seen in the figure below:



II a. to *appoint* a person to a position, office, &c. :—God gescifte ænne swā gerādne mann þe āhte gewæld ealles ðæs splottes God *appointed a man of this kind to be the owner of all the plot*, Hml. S. 23, 414. Beón gescyfte (*ordinentur*) gebrōþru, þā tō sealmsange geæmtian, Angl. xiii. 444, 1121. Gescifte, 422, 815. Sýn twēgen tō þām sylfan gescyfte (*destinati*), 410, 640. III. to *arrange, regulate* :—God sylf ðæs þingc swā gescifte and mid his fadunge gedihte, Hml. S. 23, 256. Hū bisceophādas wurdan ærest āstealde and be Godes dihte mannum gescyfte, Wlfst. 176, 7. II a. to *regulate a measure, weight, &c.* :—

Figure 9 – Sense Numbering Error in “ge-sciftan” (Bosworth and Toller, 1921, p. 403)

As can be seen, the original sense hierarchy is IIa > III > IIa, which is an apparent error. Looking at the senses, it is clear that the definition “to regulate a measure, weight” is a sub-sense of the preceding definition “to arrange, regulate”, therefore, this relationship has to be reflected in the sense numberings, leaving us with the corrected hierarchy IIa > III > IIIa which is then used in the web application. Other than that, the actual text of BT is left unchanged<sup>25</sup> – some may argue that leaving out the encyclopaedic comments may be more in line with the current lexicographic practice or that it may be more user-friendly to abridge some of the convoluted entries, yet, this would disrupt the balance of being true-to-original and user-friendly at the same time as only one of these notions (the user-friendliness) would be applied.

<sup>25</sup> Except for the addition of basic grammatical properties such as parts-of-speech and verb categorization. These are however, displayed in a section separate to the main body of the entry (see Figure 5, top left corner)

When it comes to user-friendliness, there have been graphical changes done that improve the readability of the dictionary, the improvement of structural readability is described in chapter 7.2 (see primarily Figure 19). In this chapter let us focus on the other improvements done. Firstly, one of the criticized notions of the original was its, already at that time, outdated orthography. What the critics meant by that was predominantly the employment of acutes instead of macrons. The web application balances the best lexicographic practice, fidelity to the original, and user-friendliness through a “toggle on” function that lets the user decide whether he wants to use the original orthography with acutes, the most lexicographically correct orthography with macrons, or a simplified orthography with no diacritics. Functions such as “toggle-on” are innately restricted to the digital medium and are a good representation of the modernization possibilities digitization brings. Another digital-only function is that of hyperlinks, which is used commonly in the web application. Instead of listing through the dictionary to find the one headword referred to in the entry you have been reading, a single click will on the web application will navigate you to your desired location. The dictionary has also become inter-connected with other sources describing OE or the later stages of English, again, through the use of hyperlinks, one can now easily navigate to further grammatical information on the OE form in Wright’s *Old English Grammar* (1914) or to information on the ME reflex in *Middle English Dictionary*.

Concerning the graphical distinctions in the text itself, the digital medium of the web application offers a wider and more user-friendly portfolio of fonts and colours than the paper medium of the original did. Where the original had to get by with a simple boldface-italics-basic font contrast for all the different lexical structures, the digitized BT employs a specialized font for each of the structures, moreover, the user-friendliness is further improved by graphical distinction of certain lexical structures not differentiated by the original such as headword form inside an OE example. Compare in the figure below taken from the already mentioned entry “dǣd-bētan”:



To make amends, give satisfaction, to be penitent, to repent; *maleficium compensare, malum bono pensāre, pœnitere*

His s̅awle w̅unda d̅ædb̅ētende gel̅ácnian  
to heal the wounds of his soul by making amends,

Homl. Th. i. 124, 14.

*To make amends, give satisfaction, to be penitent, to repent; maleficium compensare, malum bono pensāre, pœnitere:—His s̅awle w̅unda d̅ædb̅ētende gel̅ácnian to heal the wounds of his soul by making amends.*

Figure 10 – Comparison of Graphical Distinctions of “d̅æd-b̅ētan” in *An Anglo-Saxon Dictionary Online* (2014, <https://bosworthtoller.com/7312>) and *An Anglo-Saxon Dictionary* (Bosworth and Toller, 1898, p. 192)

In the web application, all the lexical structures are given a unique font – blue boldface for PDE translational equivalents, blue italics for Latin equivalents, basic font for OE examples, and red font for translations with the extra distinction of the headword form by boldface and underlining and thus making the navigation for the user much simpler.

This chapter served as a description of the underlying idea of BT digitization – to attain a balance between being true-to-original whilst still offering a user-friendly modern tool for researchers and the public alike. This was achieved by making as few changes to the original text as possible, with no omissions from the original text and changes to it being made very sparsely only in case of apparent errors. The modernization and increased user-friendliness springs from making use of the various features that the digital medium, in comparison to paper medium, offers, including the use of hyperlinks or toggle functions. Now, let us turn to a more in-depth description of the structures found in BT and the ways in which they are marked in the XML programming language and further transformed into the web application.

## 7. Formats of the Bosworth-Toller Dictionary

The following chapters will serve as an in-depth description of the various formats of BT. Firstly, the focus will be given to the microstructure of the source text, i.e. the original printed version. Heed will be paid to the various structures and their graphical distinctions that appear consistently throughout the dictionary (as opposed to marginal occurrences studied later in the paper). Secondly, the current format derived from the printed original through the means of digitization will be described. The current format consists of the current XML markup validated against a custom schema generated for the purposes of BT. The focal point will be the XML markup, how it reflects the structures of the original, and whether it facilitates a user-friendly and truthful-to-original HTML transformation. The third format of the dictionary is the web application which is derived from the current format through XSLT. This format will not have a dedicated chapter, rather, it will be referenced throughout other chapters (mainly the current format chapter) as the (im)possibility of certain graphical distinctions in the web application is based on the (non)existence of a particular markup in the current XML format. The last format described will be the novelty this paper brings to the formats of BT – the TEI-Lex 0 format. This format, unlike the current format, uses standardized markup validated against a pre-made schema that ensures interoperability between various texts following this schema. The TEI-Lex 0 format can be partially derived from the current format by a 1:1 conversion (done through XSLT) from the current format elements to TEI-Lex 0 elements. At other times, in order to have a valid TEI-Lex 0 document, some new elements have to be added to the current format so that the conversion is possible. This chapter will compare the two formats, describe the common ground they share, and find solutions to parts that may prove difficult to converse. A simplified diagram of the BT formats can be found below:

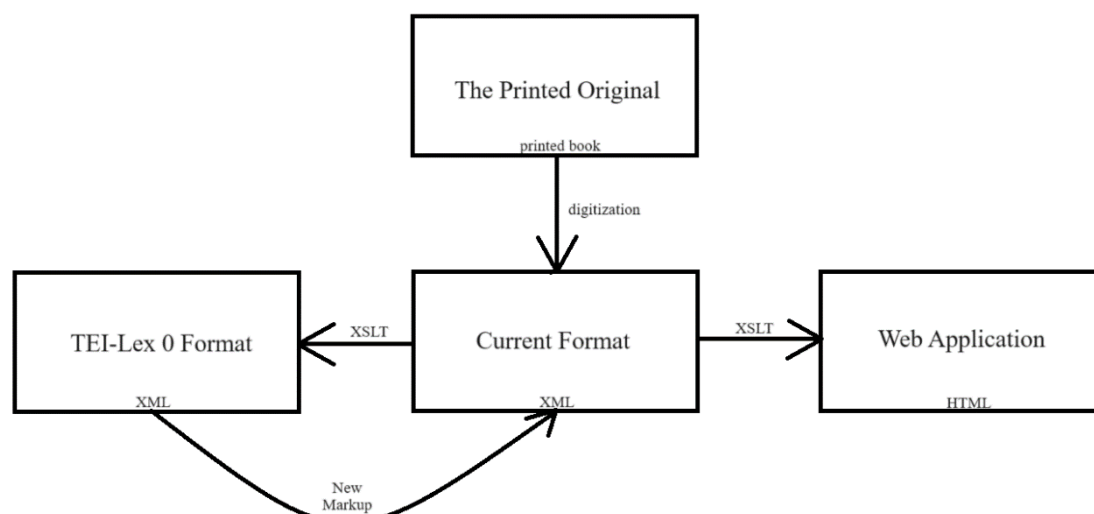


Figure 11 – Formats of the Bosworth-Toller Dictionary

## 7.1 Format of the Printed Version

The printed version, also referred to as “the original” throughout this paper is the template after which all the other formats are modelled. It can be divided into three main parts, which all contain slight structural differences<sup>26</sup> – Bosworth’s section of the main volume, Toller’s section of the main volume, and Toller’s supplement. The main focus of this chapter is to explore the consistent structural spine of BT, describe the categories it consists of, and illustrate the graphic means by which the particular categories are distinguished. For the purposes of this, a largely consistent and prototypical entry from the dictionary was chosen. It also has to be mentioned that an entry typically contains non-recurring categories such as the definition or grammatical variants category and recurring categories such as the example category. For the sake of conciseness, only a singular occurrence of a recurring category is described; regarding non-recurring categories, all of them will be illustrated. Below see the chosen entry “stingan” in its original form with described occurrences underlined in blue and omitted recurring occurrences non-underlined:

---

<sup>26</sup> E.g. different uses of parentheses in sections by Bosworth and Toller respectively, adding editorial information in the supplement, or presence of derived forms category in section by Bosworth – notable differences are described further into the paper.

**stingan**; *þ. stang, pl. stungon; þþ. stungen.* **I. to thrust something into:**—*Sting ðin seax on ða wyrte, Lchdm. ii. 346, 12. Stingaþ stranglic sār on his eāgan, Wulfst. 141, 4. Nim ān federe, and styng on hys mūde, Lchdm. iii. 130, 17. Wæs on slæpe ætywed ðæt hyre man stunge āne syle on ðone bōsum, Shrn. 149, 1. Crist hēt stingan sweord in scæde, Charter quoted by Lye.* **I a. fig. to thrust one's self into the affairs of another, to exercise authority.** *v. in-, on-sting:*—*Nā stinge nān mann on ðæt land, būton se hýred æt Xþes cyrcean, Chart. Th. 578, 6. Ic habbe ðæt geleornod, ðæt nān læwede man nāh mid rihte tō stingan hine on āre cirican, nā an ān ðara ðinga ðe tō cyrcan belimþþ. And for ði wē forbeodaþ eallan læwedan mannum æure ænne hlānordscipe ouer cyrcan, Cod. Dip. B. i. 137, 24. (Cf. *Icel. þū hefir mjök stungizk til þessa mǫls thou hast meddled much with this case.*) **II. to prick with something, to sting, stab, pierce:**—*Swā swā seó beó sceal losian, ðonne heó hwæt vringa stingþ. Bt. 31, 2; Fox 112, 26. Stingeþ, Met. 18, 7. [Wyrn] stingeþ niēten, Salm. Kmbl. 308; Sal. 153. Hē mid gāre stang wlančne wicing, Byrht. Th. 135, 55; By. 138. Stinge *transfigat*, Anglia xiii. 37, 276. Gif þorn stinge man on fōt, Lchdm. ii. 336, 20. Gif hine beón stingen, iii. 168, 13. Se læce his seax hwæt, ærdonðe hē stingan wille, Past. 26; Swt. 187, 6. Se cāsere hine hēt stingan mid irenum gyrðum, Shrn. 115, 24. Stingaþ hyne mid sāre on his eāgan, L. E. I. prm.; Th. ii. 398, 19. [*Goth. us-stiggan to thrust out; Icel. stinga to sting, stick, stab.*] *v. ā-, be-, ge-, of-, on-, tō-, þurh-, under-stingan.***

Figure 12 – Structures Studied in “stingan” (Bosworth and Toller, 1898, p. 921)

As can be seen above, we are given the headword in boldface followed by its inflected variants with grammatical information (signified by italics) in which these variants arise. This head category is separated from the following sense category by a long blank space. The sense category contains the numbering<sup>27</sup> and the definition in which the translational equivalent is expressed through italics and explanation through basic font. The following symbol “: —” separates the sense category from the following examples category. This category contains singular example quotes in Old English, optionally followed by a Latin or PDE translation written in italics and a mandatory reference to the manuscript(s) from which the quote (or translation) was taken. Each example is followed by a short blank space to distinguish the end of one example from the beginning of another. Most frequently, in a multi-sense entry, such as our illustration, the examples category would be followed by a long blank space and the next sense category in which the above-mentioned categories would recur. Yet, optionally, several other categories may follow before the commencement of the next sense category, in that case,

<sup>27</sup> The Roman numeral signifies a “super-sense” in which „sub-senses” (signified by Latin alphabet or Arabic numerals) are nested, e.g. sense I a. is a sense dependent upon sense I, whilst sense II is not.

the optional categories are related only to this particular sense and not to the whole entry. These can be comparisons introduced by the “cf.” abbreviation (see the illustration above before sense II.), related entries signified by the label “v.” (see), or etymological information in the square brackets “[ ]”. For such occurrences, please see the figures below:

**I c. in**

*reference to the heavenly bodies, tō setle gān, etc. (cf. Fr. le coucher du soleil, le soleil se couche) to set:—*Syððan sunne beó on setle *after sunset*, Lchdm. iii. 8, 19. Ðonne heó (*the sun*) tō setle gæþ, Bt. 39, 3; Fox 214, 27; Salm. Kmbl. 186, 6. Ðā ðā sunne eode tō setl *cum occubisset sol*, Gen. 15, 17. Ær sunne tō setle eode *usque ad occasum solis*, Ex. 17, 12. Ðā sunne tō setle eode *cum occidisset sol*, Mk. Skt. 1, 32. Sunne sáh tō setle, Chr. 937; Erl. 112, 17. Ðonne heó (*the sun*) on setl eode, Bt. 5, 23; S. 645, 26. Ðonne hió on setl glídeþ, Met. 28, 39. Se æfenstiorra on setl glídeþ, 29, 27, 31. On setel, Salm. Kmbl. 202, 34. v. setl-gang. **II. a seat, place where one abides, an abode,**

Figure 13 – Related Entries Structure in “setla” (Ibid., p. 866-867)

**14589.]** **I a. to sink** as the sun to its setting:—Heó (*the sun*) sính tō ðam tǣcne (*Aries*) óþ æfen, Anglia viii. 307, 20. Tungla torhtast tō sete sígeþ, Menol. Fox 221; Men. 112. Ealle stiorran sígaþ æfter sunnan under eorþan grund, Met. 29, 15. Sió æþele gesceaft (*the sun*) sáh tō setle, Chr. 937; Erl. 112, 17. [The sunne arist anes a dai and eft siged, O. E. Homl. ii. 109, 22.] **I b. in a figurative sense:—**

Figure 14 – Etymology Structure in “sigan” (Ibid., p. 872)

The same categories can also be found at the end of an entry, i.e. following the last examples category (of the last sense “super-category”) as can be seen in the illustrative entry (Figure 12) with “[*Goth. us-stiggan ...*] v. á-, ...” after the conclusion of examples category in sense II. In this case, these categories pertain to the whole entry and not just the last-mentioned sense (exceptions may exist, see “sténan” (Ibid., 908)). The last two categories not mentioned yet, are the derivational forms category, which is optional and can be found only at the end of an entry as the last category before the commencement of the next entry, and the editorial information category which is found at the opposite end of the microstructure, that is, as the first category after the headword:



**II. bishops were sometimes subject to an abbot, as they were to the abbots of Iona:—**Nū, sceal beón æfre on Iī abbod, and nā biscop; and ðan sculon beón underþeódde ealle Scotta biscopas, forðan ðe Columba [MS. Columban] was abbod, nā biscop now, in Iī [Iona], there must ever be an abbot, not a bishop; and to him must all bishops of the Scots be subject, because Columba was an abbot, not a bishop, Chr. 565; Th. 32, 10–16, col. 1. [Laym. abbed: O. Frs. abbete: N. Ger. abt: O. H. Ger. abbat: Lat. abbās; gen. abbātis an abbot: Goth. abba: Syr. אבא abba father, from Heb. אב father, pl. אבות abot fathers.] **DER. abbad-dōm, -hād, -isse, -rice: abboda.**

**abbad-dōm an abbacy. v. abbud-dōm.**

Figure 15 – Derivation Structure in “abbad” (Ibid., p. 2)

**ge-fīperhamod. Add: v. fīper-hama.**  
**ge-fīperian. Add: Gefīperede pennata, Ps. L. 77, 27; Bl. Gl. Gefīderadra pennatorum, Kent. Gl. 2.**  
**ge-flāschamod. Add:—Geflāschamod incarnatum, An. Ox. 944. v. flāschama; ge-flāscod.**  
**ge-flāscness. Add:—Ic hālsie ðē þurh ūres Drihtnes geflāscnyse, Ll. Lbmn. 415, 11.**

Figure 16 – Editorial information in Supplement (Bosworth and Toller, 1921, p. 327)

The derivational forms category signaled by the abbreviation in capital letters “DER.” is quite a rare category in the BT and is mostly used only in the beginning of the dictionary written by Bosworth. Its position in the microstructure is the same as of the category of related entries (see last lines of Figure 12) and whilst the content of these categories is not exactly the same, it often largely overlaps – a derived form of the headword is frequently the lemma of a related entry. Hence the rarity of the derived forms category in the later parts of the dictionary, where its place is taken by the related entries category. Figure 16 shows the category of editorial information and based on the illustration provided it may seem to be a part of all entries. However, taking in consideration the total number of entries it can also be said to be uncommon as it is used very rarely in the main volume of the BT and is a category almost exclusive to the supplement of the dictionary from which the figure was taken. This category can be expressed through three imperatives “Add, Dele, or Subst. signaling whether something needs to be added, deleted, or substituted in the main volume entry.

The categories shown above are all the categories that BT offers and in an ideal world with no exceptions or mistakes, all entries should adhere to this entry microstructure. Every category is neatly distinguished and the hierarchy between particular categories is clear. A hypothetical

schema for a “perfect” BT would then look like this with the categories in brackets being optional:

1. Headword
2. Orthographic variants and inflected forms + grammatical information
3. Definition
4. (Sense marker)
5. (Definition for the particular sense)
6. Example(s) in Old English
7. (English or Latin translation(s))
8. Reference(s)
9. (Etymological information)
10. (Related entries)
11. (Derived forms)

Due to the constraints of the paper medium, the distinctions between the various categories are made up of only font differences inside a particular category (e.g. italics for translation equivalents and base font for explanations in the definition category) and blank spaces or metamarks between two different categories (e.g. the blank space between example and etymology category). Such a structure can be considered linear and difficult to navigate, a problem that is resolved by the embedding structure of XML and its further transformation to the HTML-based web application which will be the focus of the following chapter.

## **7.2 Current XML format**

The current XML document is validated against a custom schema designed for the specific needs of the BT, its foundations were laid by Ondřej Tichý (2007), the leader of the digitization process, with improvements being made as time progressed and new elements were being added. Before delving into the close description of the current XML hierarchy, a clarification of what the purpose of this document is will be given. The three main principles of the current schema are ease of editing, terminological accuracy, and the possibility of user-oriented XSL transformation.

Ease of editing is achieved through a largely non-constraining schema (a small number of constraints reflects the inconsistent nature of the printed version), utilizing only the elements needed for a full-fledged user experience of the web application, i.e. there is no need to tag categories that would not improve the user experience such as metamarks or punctuation. Terminological accuracy reflects the fact that the current XML document uses elements in accordance with the actual categories found in the printed version, i.e. instead of utilizing elements such as “<italics>” for a text in italics in the web application, terminologically correct “<equiv>” (short for translational equivalent) is used. This allows for a clearer interpretation of

the XML document, as a distinction between two different elements surfacing with the same typography is retained. This furthermore grants the possibility for a less laborious conversion to a TEI-conformant XML document. Lastly, the possibility of user-oriented XSL transformation, i.e. implementation of the web application based on the XML data, is attained through an embedding “parent-child” structure (see end of this chapter) that is conducive to the clear and user-friendly interface of the web application. For an illustration of the current XML format using the same excerpt from the entry “stingan” as previously (see Figure 12) see below:

```

<entry id="029006">
  <form>
    <orth>stingan</orth>
    <search>stingan</search>
    <sort>stingan</sort>
  </form>
  <gramGrp>
    <pos>verb</pos>
    <subc>strong</subc>
  </gramGrp>
  <column name="body">
    <grammar>p. <infl func="p."><var>stang</var></infl>
    <sense num="I">
      <snum>I.</snum>
      <def><equiv lang="eng">to thrust</equiv> something into</def>
      <examples>
        <ex><oe><cit>Sting</cit> ðín seax on ða wyrte.</oe>
      </References><ref>Lchdm. ii. 346, 12.</ref></References>
      </ex> </examples>
    </sense>
    <sense num="Ia"> <snum>Ia</snum>
      <def> fig. <equiv lang="eng"> to thrust</equiv> one's self into the affairs of
      another</def>
      <see>v. <a href="020489">in-</a></see>
      <examples><ex><oe>Ná <cit>stinge</cit> nán mann on ðæt land, búton
      se hýred æt Xp-es cyrcean.</oe>
      <trans lang="eng">to exercise authority.</trans>
      <references><ref>Chart. Th. 578, 6.</ref></references></ex>
      </examples>
      <comment>(Cf. Icel. Þú hefir mjök stungizk til þessa máls
      thou hast meddled much with this case.)</comment> </sense>
    <sense num="II"> <snum>II</snum>
      <def> <equiv lang="eng">to prick with something</equiv>, <equiv lang="eng"> to
      sting</equiv>, <equiv lang="eng">stab</equiv>, <equiv
      lang="eng">pierce</equiv> </def>
      <etym>[<item><source>Goth.</source> <cog>us-stiggan</cog> <equiv
      lang="eng">to thrust out</equiv></item>:
      <item><source>Icel.</source> <cog>stinga</cog></item>]</etym></entry>

```

Figure 17 – Current XML Structure of “stingan”

To clarify the terminology used for the description of XML documents, the texts in chevrons will be referred to as elements. Elements can be either bare, using only the blue colour for the name of the element, or they can be complex, where orange stands for the attribute (of an



element), and red signifies the value of the attribute. Text in black, that is, outside of chevrons, is the actual text of the dictionary. A slash stands for an element's ending. For instance, “<equiv lang="eng">to thrust</equiv>” is an element “equiv” (translation equivalent) with the attribute “lang” (language), whose value is “eng” (English). The content of such element is an English translational equivalent of the given headword, in our case “to thrust” being the equivalent of “stingan”.

Turning to the entry, it begins with several elements (referred to as categories in the printed version) that are not expressed in the printed version. Firstly, the element <entry id> is added for macro-structure clarity and referential reasons – every headword is given a unique identification number which simplifies the referential process in the case of homonymous headwords. The <form> element contains the elements <orth> which is synonymous with the headword category from the previous chapter, <search> for simpler access to the entry in the web application as diacritics unknown to PDE are removed, and lastly <sort> that allows for a list of all the entries in the specific alphabetical order of OE<sup>28</sup> and this, in return, grants easier access to the complementary entries in the supplement<sup>29</sup> of BT as can be seen below where main body entry “strégan” and supplement entry “strégan” are “closer” to each other due to the alphabetical list:

---

<sup>28</sup> OE alphabet included graphemes obsolete in PDE such as thorn (þ) or eth (ð)

<sup>29</sup> In the printed version, the same headword in the main body and the supplement is located far away from each other, making it hard to navigate. The <sort> element solves this issue and enhances the user-experience.

Figure 18 – Connectedness of Main Volume and Supplement Entries, exemplified on “strégan” (Anglo-Saxon Dictionary Online, 2014, <https://bosworthtoller.com/29143>)

The next element is <GramGrp> and it contains the grammatical information concerning the headword. Firstly, the <pos> (parts of speech) element which is part of all entries in the XML document, yet in the printed version of BT is unexpressed for nouns and verbs which can be easily distinguished based on the context (other parts of speech are expressed even in the printed version). For verbs, there is also the element <subc> which specifies the membership to either weak or strong class, this element, again, has no category counterpart in the printed version and has to be deduced based on context or OE grammars. For nouns, there is the element of <gen> (grammatical gender) which, on the other hand, is expressed in the printed version.

The elements following <GramGrp> in the XML document already find their direct counterparts in the original. The original structure with its parent<sup>30</sup> element counterparts in the current format can be seen below:

1. Headword
2. Variants + inflected forms = <grammar>
3. Definition = <def>
4. (Sense marker) = <sense>
5. (Definition for the particular sense) = <def>
6. Example in Old English = <ex>
7. (English or Latin translations) = <trans>
8. Reference(s) = <references>
9. (Etymological information) = <etym>
10. (Related entries) = <see>
11. (Derived forms) = <der>

<sup>30</sup> Parent element = element in which “child” elements are nested. Parent elements are hierarchically superordinate to child elements.

Child elements, on the other hand, do not reflect categories, but rather special occurrences inside of a category. These can be either typographically expressed in the printed version such as `<infl>` (expressed through italics in the printed version) and `<equiv>` (also italics) or unexpressed, such as the element `<cit>` marking the use of the headword in each example (signified through boldface in the web application for better user-experience). As mentioned at the beginning of the chapter, these elements are good not only for the possibility of a digitized version faithful to the original (occurrences typographically distinct in the original are typographically distinct in the web application) or for a better readability and user experience as in the case of `<cit>` which is unexpressed in the original, but also, importantly, to uncover the metastructure of the dictionary and correctly mark it up.

One of the notions mentioned in the introduction of this chapter is the parent-child structure of the current XML document. This structure allows for a user-oriented XSL transformation as each parent element/category is clearly distinguished from the rest. Where the printed version relied on metamarks and blank spaces, not always being clear where one category ends and another starts, the parent elements in XML (and hence the web application) draw clear lines. The parent elements are then taken as separate objects with their content – child elements – as illustrated here:

# CEÓ

Noun [ feminine ]

Dictionary links



Grammar

CEÓ, *ció*; indecl. f.

Wright's OE grammar

§405:

I. A **CHOUGH**, a bird of the genus *corvus*, *ajay*, *crow*, *jackdaw*; *cornix*, *gracculus*, *monedula*

Show examples

Etymology

[Scot. *keaw*: ♦ Dut. *kauw*, *f.* ♦ M. H. Ger. *kouch*, *m.*  
a horned owl: ♦ O. H. Ger. *kaha*, *f.* ♦ Dan. *kaa*,  
*kaje*, *m. f.* ♦ Swed. *kaja*, *f.* ♦ Icel. *kjóí*, *m.* a sea-  
bird.]

Linked entries

v. *chýae*, *ció*.

Figure 19 – Graphical Distinctions of Parent Elements in the Web Application, exemplified on “ceó” (Ibid., <https://bosworthtoller.com/6032>)

In Figure 19, four parent-child structures are marked. The green frame reflects the structure of <grammar> (the parent element) and its children <infl> and <var>, the dotted frame shows the structure of parent element <sense> with its children <def> and <examples><sup>31</sup>, the red frame is the parent element <etym> consisting of child elements <item>, and lastly the blue frame shows the transformation of the XML parent element <see> with its children tagged by <a href>. Therefore, it can be concluded that the linear structure of the printed version is via the current XML markup and its transformation made into a more engaging and readable structure with clearly defined vertical hierarchy which at the same time remains faithful to the original. Compare the current XML hierarchy below (parent elements are highlighted by the colours they were assigned in the previous figure with purple substituting the dotted frame and orange for the parent element <der>) to the more linear structure of the printed version (see end of chapter 7.1):

<sup>31</sup> <examples> being at the same time the parent for elements <ex>, <trans>, and <References>. <References> in turn is the parent element for <ref>. The hierarchy works exactly as in family related kinship terms, e.g. <Sense> is a great-great-great-parent of <ref> (<Sense> is four levels higher than <ref>). For simplicity's sake, this paper works only with the terms parent and child however big the level discrepancy is.

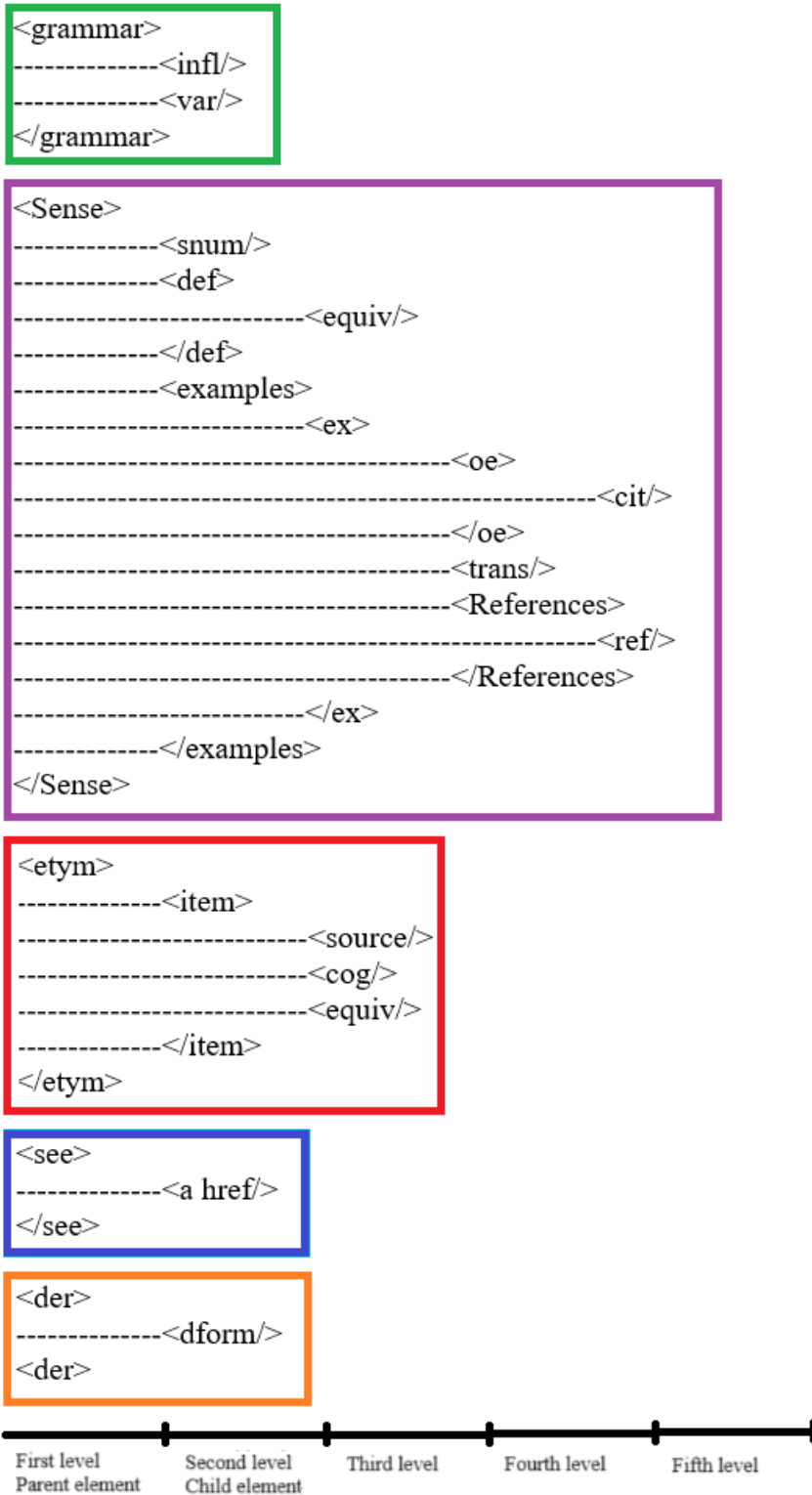


Figure 20 – Universal Current XML Structure

To summarize, the current XML format is modelled after the original version of BT. It takes all the structures and adds some new ones that improve the precision of the markup which in turn improves the readability of the web application. The less engaging linear structure of the original caused by the inherent constraints of the paper medium is transformed into a more

reader-centered vertical structure with clearly defined items (child elements) and categories (parent elements) to which they pertain. This structure in turn facilitates an engaging user-friendly web application as seen in Figure 19. Whilst the current format is focused on the balance between the most economical way and the best possible user-friendliness (whilst staying true to the original), the format of the following chapter is not so much interested in these notions. TEI-Lex 0 formatting is mostly about being as precise as possible, notwithstanding the economicity of the task. The final product is thus not oriented towards the general public as much as it is to researchers and lexicographers looking for the maximalist version of the given dictionary. The unquestionable advantage of this way of formatting is the fact that it is standardized, which guarantees preservation and inter-operability with other digitized dictionaries following the same standard.

### **7.3 TEI Lex-0 XML Format**

A TEI Lex-0 XML document is validated against the standard TEI Lex-0 schema created by the TEI Lex-0 team (Tasovac, Romary et al. sec.13.3) with the full documentation being available on the TEI Lex-0 website (sec.12). The purpose of this type of formatting is to have a digitized version of BT that is completely transparent to anyone knowledgeable of the TEI Lex-0 standard, and therefore ensuring the preservation of the dictionary whilst making it inter-operable with all other TEI Lex-0 formatted dictionaries. As the aims of this way of XML formatting are different from the ones mentioned in the previous chapter, the elements and their hierarchies are changed. To give two examples, the parent-child hierarchy useful for the interface of the web application is lost in certain structures, as this way of tagging is not in line with TEI Lex-0 guidelines, on the other hand, a sizeable number of new elements have been added as the TEI Lex-0 requires a much more precise tagging, even of elements that may seem inconsequential, e.g. the element `<pc>` to mark up punctuation marks that are not already embedded in another element. The purpose of this chapter is to compare the current format and TEI-Lex 0 format, find structures common for both the formats and come up with solutions where the difference in structures would not make a conversion possible. It has to be kept in mind that as TEI-Lex 0 is the more markup-heavy format but at the same time a format that will be derived from the current format, the solutions given would take many added elements to the current format and re-editing of the majority of the dictionary. The examples are thus to be taken as the best-case scenario, but in reality (at least in the first stages of the conversion) the TEI-Lex 0 format will not be as precise as it should be, simply due to the fact that the current format does not facilitate such a conversion. This will not be a sizeable problem regarding this

chapter but once we get into more complex and non-prototypical structures in chapter 8, this fact will play a huge role. But before we delve into the intricacies of the TEI Lex-0 XML formatting, a preliminary step, required for all TEI-conformant documents, has to be taken.

This step is the filling out of the <teiHeader> element, where all the basic information concerning the digitized version of the dictionary is kept. For the sake of conciseness, only a few examples, upon which the function of <teiHeader> will be shown, are given below<sup>32</sup>:

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI type="lex-0">
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>An Anglo-Saxon Dictionary</title>
    </titleStmt>
  </fileDesc>
  <profileDesc>
    <langUsage>
      <language ident="ang" role="sourceLanguage">Old English</language>
      <language ident="la" role="targetLanguage">Latin</language>
      <language ident="en" role="targetLanguage">English</language>
    </langUsage>
  </profileDesc>
</teiHeader>
</TEI>
```

Figure 21 – TeiHeader Abridged

In the figure, there are two child elements of <teiHeader>, that is <fileDesc> and <profileDesc>, both being required elements of the <teiHeader>. The other child elements <sourceDesc>, <encodingDesc>, and <revisionDesc> (not present in the figure) are not mandatory but are highly recommended (Ibid.). The element <fileDesc> contains the most information with up to six child elements out of which each permits a number of further child elements. This structure is conveyed in the figure by the hierarchy of <fileDesc> with its child <titleStmt> which in turn has the child <title> that contains the actual title of the BT. The second element portrayed is the <profileDesc> which is notably less convoluted than its predecessor as it permits only a single child <langUsage> which requires at least a single child <language>. The element <language> has two mandatory attributes, “ident” whose value is the ISO 639 code of the given language<sup>33</sup>, and “role” whose value is determined by the scope of use in the given dictionary, the possible values are source-, target-, object-, and workingLanguage.

<sup>32</sup> For a full overview of the required and optional elements embedded in the <teiHeader> element see Tasovac, Romary et al. (sec.2)

<sup>33</sup> For living languages ISO 639-1 is used and extinct languages (such as Old English) use the ISO 693-2 code.

With validly (and truthfully) filled out <teiHeader> it is possible to now turn to the core of the TEI Lex-0 formatting which is the entry structure. This formatting will be shown, again, on the abridged version of the entry “stingan” (cf. Figures 12 and 17):

```
<entry xml:id="_29006" xml:lang="ang">
  <form type="lemma">
    <orth>stingan</orth>
  </form>
  <gramGrp>
    <gram type="pos">verb</gram>
    <gram type="inflectionType">strong</gram>
  </gramGrp>
  <form type="inflected">
    <gram type="p.">p.</gram> <orth>stang</orth></form>
  <sense xml:id="_29006.I">
    <num>I.</num>
    <def><cit type="translationEquivalent">to thrust</cit> something into</def>
    <cit type="example" xml:lang="ang"> <quote><ref type="oRef">Sting</ref> ðin seax
    on ða wyrte.</quote>
    <listBibl><bibl>Lchdm. ii. 346, 12</bibl><pc>.</pc></listBibl></cit>
    <sense xml:id="_29006.I.a">
      <num>Ia.</num> <usg type="meaningType">fig.</usg>
      <def><cit type="translationEquivalent" xml:lang="en">
        <orth>to thrust</orth></cit> one's self into the affairs of another,</def>
      <xr type="related"> <lbl>v.</lbl>
        <ref type="entry" target="_20689">in</ref>
        <pc>.</pc>
        <ref type="entry" target="_24767">on-sting</ref></xr>
      <cit type="example" xml:lang="ang"><quote>Ná <ref type="oRef">stinge</ref> nán mann on
      ðæt land, búton se hýred æt Xp̄es cyrcean.</quote>
      <cit type="translation" xml:lang="en"><quote>to exercise authority.</quote></cit>
      <listBibl><bibl>Chart. Th. 578, 6</bibl><pc>.</pc></listBibl></cit>
      <note>(Cf. Icel. Þú hefir mjök stungizk til þessa máls thou hast meddled much with this case.)</note>
    </sense></sense>
    <sense xml:id="_29006.II">
      <num>II.</num>
      <def><cit type="translationEquivalent" xml:lang="en"><orth>to prick</orth></cit> with something,
      <cit type="translationEquivalent" xml:lang="en"><orth>to sting</orth></cit><pc>.</pc>
      <cit type="translationEquivalent" xml:lang="en"><orth>stab</orth></cit><pc>.</pc>
      <cit type="translationEquivalent" xml:lang="en"><orth>pierce</orth></cit><pc>.</pc></def>
      <cit type="example" xml:lang="ang">
      <quote>Swá swá seó beó sceal losian, ðonne heó hwæt yrringa <ref type="oRef">stingþ</ref>.</quote>
      <listBibl><bibl>Bt. 31, 2</bibl><pc>.</pc> <bibl>Fox 112, 26</bibl><pc>.</pc></listBibl></cit></sense>
      <etym><metamark>[</metamark>
        <cit type="cognate">
          <lang>Goth.</lang>
          <form> <orth xml:lang="got">us-stiggan</orth></form>
          <cit type="translationEquivalent" xml:lang="en">
            <orth>to thrust out</orth></cit> </cit> <pc>:</pc>
        <metamark>]</metamark></etym>
    </sense>
  </entry>
```

Figure 22 – Tei-Lex 0 XML structure of “stingan” (Bosworth and Toller, 1898, p. 921)

Before commencing with the description of the structure, it is important to mention that this XML formatting is the more convoluted of the two (cf. Figure 17), there are more elements (if we discount the parent elements such as <examples> in the previous format) and many more mandatory attributes. This is due to the nature of TEI-Lex 0 underlying idea of marking all items that carry meaning with the exceptions of spaces, i.e. even things that would probably not be graphically distinguished from the rest of the text in a theoretical web application as, for example, punctuation marks, require their distinct markup in TEI Lex-0 formatting. As this way



of XML formatting is a novelty for the BT, a more in-depth description with direct comparisons to the current format will be given, starting with the “pre-definition” part<sup>34</sup>:

```

<entry id="029006">
  <form>
    <orth>stingan</orth>
    <search>stingan</search>
    <sort>stingan</sort>
  </form>
  <gramGrp>
    <pos>verb</pos>
    <subc>strong</subc>
  </gramGrp>
  <column name="body">
  <grammar>p. <infl func="p."><var>stang</var>
  </infl></grammar>

```

---

```

<entry xml:id="_29006" xml:lang="ang">
  <form type="lemma">
    <orth>stingan</orth>
  </form>
  <gramGrp>
    <gram type="pos">verb</gram>
    <gram type="inflectionType">strong</gram>
  </gramGrp>
  <form type="inflected">
    <gram type="p.">p.</gram> <orth>stang</orth></form>

```

Figure 23 – Comparison of “pre-definition” Structure in Current XML Format and Tei-Lex 0 Format, exemplified on “stingan” (Ibid.)

For the most part, this section of the structure is rather similar, some attributes required by the TEI Lex-0 schema were added such as the source language of the entry or the type of form, on the other hand, some elements needed for the web application (as the TEI Lex-0 is meant for reading in XML rather than on the web) such as <search> and <sort> were lost. The only notable change in this part is the omission of the element <grammar> from the current format. As mentioned, the parent-child hierarchy used so often in the current format is for some structures deprecated by the TEI Lex-0 and so the element <grammar> is lost. Also, in the current format, the information on the type of the form is given as an element <var> (for orthographical variants) or combination of <var> and <infl> (for inflected forms), whereas in the TEI-Lex 0 this information is conveyed as an attribute to the element <form> that envelopes both the orthographical form of the word and its grammatical properties signified by <gram>. The

<sup>34</sup> For this and all subsequent figures in this chapter, the current XML format will be at the top and the TEI Lex-0 format at the bottom, divided by a red line.

current format takes a slightly different way of achieving the same message by the attribute “func” on the element <infl>. All in all, the pre-definition part of both the formats is similar and should not pose many problems in conversion from the current format to the TEI Lex-0 format as the element <gram><sup>35</sup> can be automatically added to the restricted set of grammatical abbreviations (“p.”, “pp.”, “pl.”, etc...) and then included in the element <form> and the elements lost in TEI-Lex 0 will be simply omitted during the transformation process.

The next section illustrated will cover the sense marking and definition of the headword:

```

<sense num="II">
  <snum>II.</snum>
  <def><equiv lang="eng">to prick</equiv> with something,
    <equiv lang="eng">to sting</equiv>,
    <equiv lang="eng">stab</equiv>,
    <equiv lang="eng">pierce</equiv>
  </def>

```

---

```

<sense xml:id="_29006.II">
  <num>II.</num>
  <def><cit type="translationEquivalent" xml:lang="en">
    <orth>to prick</orth></cit> with something<pc>,</pc>
    <cit type="translationEquivalent" xml:lang="en">
    <orth>to sting</orth></cit><pc>,</pc>
    <cit type="translationEquivalent" xml:lang="en">
    <orth>stab</orth></cit><pc>,</pc>
    <cit type="translationEquivalent" xml:lang="en">
    <orth>pierce</orth></cit><pc>,</pc></def>

```

Figure 24 – Comparison of Sense and Definition Structures

Starting with the attribute on the element <sense>, it can be seen that where the current XML uses an attribute “num” with a non-unique value (in this case “II”), the TEI Lex-0 prioritizes attribute “xml:id” which in turn requires a unique value, which is made of the entry id and the sense number divided by a dot. Regarding the actual translational equivalents and explanations, the markup is the same for both of the formats. The only other difference is the tagging of punctuation marks which, again, can be simply automated during the transformational process. Two other elements may be used in this category that are not used in Figure 24, these are <gloss> used for additional information, often in parentheses:

<sup>35</sup> The attribute “type” shown in the figure is not mandatory

**stirn-lic; adj.** I. *hard, harsh*:—Warna ðæt ðú nán þing styrn-lices ne sprece ongēn Iacob cave, ne loquaris contra Iacob quidquam durius, Gen. 31, 29. II. *hard, unpleasant, severe (of weather)*:—Hwiltidum ðeós woruld is gesundful and myrige on tō wunigenne, hwilon

Figure 25 – Definitional Gloss, exemplified on “stirn-lic” (Bosworth and Toller, 1898, p. 922.)

The underlined phrase would then be tagged as “<gloss> (of weather) </gloss>” in a TEI Lex-0 document. The second element is <usg> which is used to denote the specific connotations of the headword such as its figurative meanings, its (in)formality, etc..., an occurrence of this can be seen in the illustrative entry in the definition of sense I a:

I a. *fig. to thrust one’s self into the affairs of another, to exercise authority.* v. in-, on-sting:—Nā stinge

Figure 26 – Usage Information in Definition, exemplified on “stingan” (Ibid., p. 921)

Which translates into TEI Lex-0 markup such as this:

```
<sense xml:id="_29006.I.a">
  <num>Ia.</num> <usg type="meaningType">fig.</usg>
```

Figure 27 – TEI-Lex 0 Markup of Usage Information in Definition

Glosses may be difficult to markup automatically as parenthesized text with other functions can be part of the definition category, however, it is worth a consideration to add it to the current format as it may improve the readability of the text in the web application. Usage on the other hand is used for a clearly restricted set of labels (fig., lit., poet., etc...) and thus would not hinder the conversion, yet, again, this element may be useful even for the web application and may be added to the current format rather than generated during the conversion to TEI-Lex 0. Other than that, the other structures are again similar, and conversion should not pose any problems.

Let us move now to the example section of the microstructure:

```
<examples><ex><oe>Nā <cit>stinge</cit> nán mann on ðæt land,
  búton se hýred æt Xp̄es cyrcean.</oe>
  <trans lang="eng">to exercise authority.</trans>
  <references><ref>Chart. Th. 578, 6.</ref></references></ex>
</examples>
```

---

```
<cit type="example" xml:lang="ang"><quote>Nā <ref type="oRef">stinge</ref>
  nán mann on ðæt land, búton se hýred æt Xp̄es cyrcean.</quote>
  <cit type="translation" xml:lang="en"><quote>to exercise authority.</quote></cit>
<listBibl><bibl>Chart. Th. 578, 6</bibl><pc>.</pc></listBibl></cit>
```

Figure 28 – Comparison of Example Structure

Once again, one of the parent-child hierarchies is lost in TEI-Lex 0 which means that during the conversion, element <examples> will have to be omitted but other than that the current format and the TEI Lex-0 format are at first glance almost identical which should result in a

simple conversion. The only problematic tag is the <quote> element which is in OE examples used in identical positions as the current element <oe> but has no counterpart in the translational section, yet it should pose no problems as the beginning of the tag is easily identifiable as following the preceding element <cit type="translation"> and the ending tag as preceding </cit>. The optional sections that may follow the examples are etymology, related entries, and derived forms (see Figure 20). The first section illustrated will be the one present at the end of the illustrative entry, that is etymology:

```

<etym>[<item><source>Goth.</source> <cog>us-stiggan</cog>
  <equiv lang="eng">to thrust out</equiv>:</item>
<item><source>Icel.</source> <cog>stinga</cog> </item>]</etym>


---


<etym><metamark>[</metamark>
<cit type="cognate">
  <lang>Goth.</lang>
  <form> <orth xml:lang="got">us-stiggan</orth></form>
  <cit type="translationEquivalent" xml:lang="en">
    <orth>to thrust out</orth></cit> </cit> <pc>:</pc>
<cit type="cognate">
  <lang>Icel.</lang>
  <form><orth xml:lang="is">stinga</orth></form></cit>
  <metamark>]</metamark></etym>

```

Figure 29 – Comparison of Etymology Structure

At the first line of the TEI Lex-0 format, there is an element with no counterpart in the current format, which is <metamark>. Meta marks are different from punctuation marks (<pc>) in that they gain a specific function for the needs of a given dictionary, in our case the symbol “[” functions as a sign of the etymology section’s beginning. As this is the only metamark in BT<sup>36</sup> and it is restricted to a single symbol following the element <etym> and the structure is otherwise identical<sup>37</sup>, the conversion will be unproblematic. However, one thing that has to be kept in mind is that TEI Lex-0 requires an attribute on the element <orth> specifying the ISO 639 language code of the cognate.

The second section that may follow the examples (or the optional etymology section) is the related entries section which can be seen also in the illustrative entry, although in a non-prototypical place<sup>38</sup>, after the definition section in sense I a:

<sup>36</sup> The only other “:---” has been omitted in the current format.

<sup>37</sup> Punctuation marks will be dealt with in the same manner as was described in the definition category.

<sup>38</sup> More about marginal occurrences in the following chapter.

```

<see>v.
  <a href="020689">in-</a>,
  <a href="024767">on-sting</a>
</see>


---


<xr type="related">
  <lbl>v.</lbl>
  <ref type="entry" target="#_20689">in-</ref>
  <pc>,</pc>
  <ref type="entry" target="#_24767">on-sting</ref>
</xr>

```

Figure 30 – Comparison of Related Entries Structure

Similarly to the etymology section, the TEI Lex-0 format begins with a new element, in this case <lbl> which is used for various abbreviations with a specific function, in terms of BT these would be mainly “v.” for see and “cf.” for compare. The only other difference is that TEI Lex-0 requires not only the target of the reference (the value of which is the unique ID of a given entry) but also the type of reference whose value is for BT’s needs either “entry” or “sense”<sup>39</sup>. The last possible section is the derived form section that will be illustrated on the excerpt of entry “abbad” (see Figure 15):

```

<der>DER.
  <dform>abbad-dóm</dform>
  <dform full="abbad-hád">-hád</dform>
</der>


---


<lbl>DER.</lbl>
<form type="derived"> abbad-dóm</form>
<form type="derived" expand="abbad-hád">-hád</form>

```

Figure 31 – Comparison of Derivation Structure

This section is again a clear showcase of the underlying similarity between the current format and the TEI Lex-0 format as without the parent element <der> they are completely identical. All in all, it can be said that for the most part, the current format and the TEI Lex-0 format are very similar, permitting the conversion from the custom elements to TEI-conformant elements. However, it has to be kept in mind that so far, it has been dealt with prototypical, structured entries and the main differences between the current and TEI Lex-0 format will be explored in the following chapter. At present time a summarizing conversion table between the formats would look as follows:

<sup>39</sup> Reference to a specific sense of an entry is possible only in TEI Lex-0 format due to the unique identification of senses (see Figure 24) but may be implemented to the current format (see “further changes”).

<entry id="uniquenumber">	<entry xml:id="_uniquenumber" xml:lang="ang">
<form>	<form>
<orth>	<orth>
<gramGrp>	<gramGrp>
<pos>	<gram type="pos">
<subc>	<gram type="inflectionType">
<gen>	<gram type="gen">
<infl>	<form type="inflected">
<infl full="fullform">	<form type="inflected" expand="fullform">
<infl func="gramfunction">	<form type="inflected"> <gram type="gramfunction">
<var>	<form type="variant">
<sense num="number">	<sense xml:id="_uniquenumber">
<def>	<def>
<equiv lang="eng">	<cit type="translationEquivalent" xml:lang="en"> <form>
<ex>	<cit type="example" xml:lang="ang">
<oe>	<quote>
<cit>	<ref type="oRef">
<trans lang="eng">	<cit type="translation" xml:lang="en"> <quote>
<references>	<listBibl>
<ref>	<bibl>
<etym>	<etym>
<item>	<cit type="cognate">
<source>	<lang>
<cog>	<cit type="cognate"> <form> <orth xml:lang="ISO639code">
<see>	<xr type="related">
<a href="uniquenumber">	<ref type="entry/sense" target="#_uniquenumber">
<dform>	<form type="derived">

Figure 32- Conversion Table Between Current XML Format and TEI-Lex 0 Format

The table above shows all the convertible elements in both the formats, yet it is important also to mention the elements that have no counterparts in their respective “counter-schema”. For the current format, these would be the parent elements: <grammar>, <examples>, and <der>. For the TEI Lex-0 format, those would be the elements marking a specific function that has no

consequence on the web application interface: <gloss>, <lbl>, <metamark>, and <usg>. Out of which only <gloss> would require manual tagging, the other elements can be easily generated during the transformation process. All in all, with a proper XML-to-XML XSLT document, it should be possible to convert most of the elements seamlessly with only a small portion of the work having to be done manually. As far as the hierarchical structure of the elements is concerned, due to the omission of some of the parent elements, some of the verticality present in the current format has been lost, however, enough has been retained to keep a clear internal structure for all the main categories (sense, example, and etym). The hierarchy is to be seen below with subordination marked by indentation (cf. current XML structure in Figure 20):

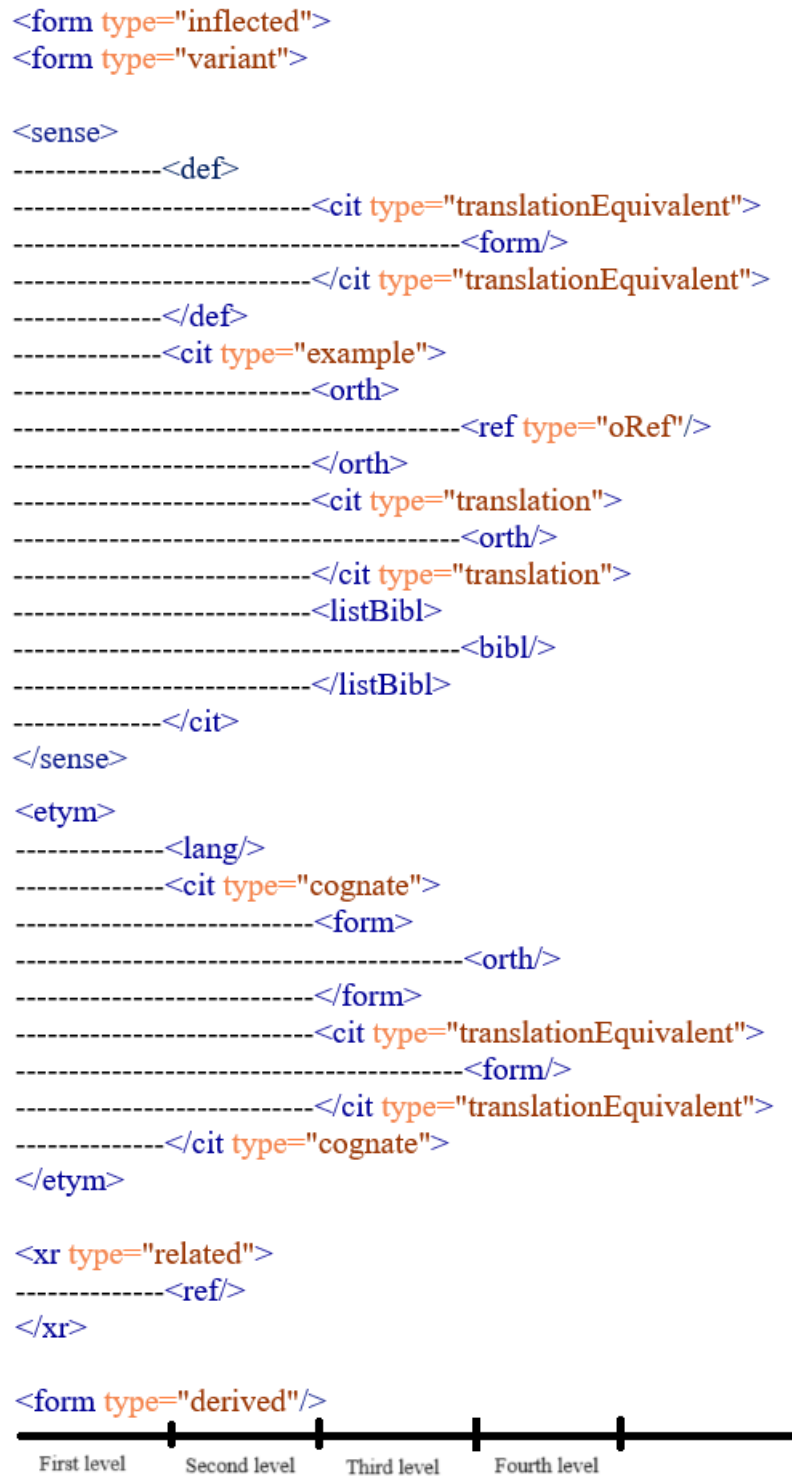


Figure 33 – Universal TEI-Lex 0 Structure

It has been shown that the underlying similarities between the current format and the TEI-Lex 0 format facilitate a simple conversion for entries that follow the prototypical structures of BT. The XSLT document will be largely based on the conversion table and the notes accompanying the automatically generated elements such as <pc> or <usg>. Unfortunately, many entries do not follow the prototypical structures, and the current format does not contain elements needed



so that a TEI-Lex 0 conversion would be valid for such marginal occurrences. Changes to the current format will be suggested – addition of new elements and attributes or standardization of markup for currently undescribed, vague structures – in order to have an improved version of the format that would, in the future, facilitate conversion to a more rigid and complete version of TEI-Lex 0.

## **8. Non-prototypical and Complex Structures**

Whereas the preceding chapters served as an in-depth description of the various formats of BT and their corresponding microstructures, the chapter at hand will focus upon occurrences that do not fully conform to these structures. As mentioned before, BT is a dictionary that is inconsistent regarding its microstructure with some inconsistencies being more regular than others. To give a comprehensive list of all the inconsistencies and their markup is a task nigh impossible and it has to be kept in mind that some structurally anomalous entries will have to be treated separately from others. However, some of the inconsistencies appear quite regularly and a standardized markup valid against the two schemas is needed. The first marginal structures described will be the parenthesized texts of BT, a distinction will be drawn between commentaries and intra-example glosses and the typologies of these structures will be given with a suggested markup for each of the type. Secondly, a closer look will be given to the etymology section where two new structures will be distinguished from the prototypical “cognate structure” and a new way of markup will be proposed. Lastly, I will give a list of further changes consisting of advocacy for small modifications to the XSLT document. These modifications, unlike the more convoluted structures in the preceding chapters, would require less manual work whilst improving the user-interface of the web application.

## 8.1 Commentaries and Notes

The largest category of parenthesized texts is made of additional information that cannot be grouped under a single element with a clear lexicographic function (e.g. <ex>, <etym>, ...), such occurrences can crop up at any level of the hierarchy as can be seen in the following figure:

**hæðen**; *adj.* HEATHEN, *pagan, gentile*; and *subst. a heathen*:—*Twā folc ðæt is Iudēisc and hæðen two peoples, that is Jew and gentile, Homl. Th. i. 206, 32. Ðes wæs hæðen hic erat samaritanus, Lk. Skt. Rush. 17, 16. Gif ungefullod cild færlīce biþ gebroht tō ðam mæssepreoste hæ hit mōt fullian sōna ðæt hit ne swelte hæðen if an unbaptized child be brought to the mass-priest suddenly, he must baptize it at once, that it die not heathen, L. Ælfc. 26; Th. ii. 352, 17; L. M. I. P. 42; Th. ii. 276, 15. Hēr sæt hæðen here on Tenet in this year a heathen [Danish] army sat in Thanet, Chr. 865; Erl. 70, 31. Óð ðone hæðenan byrgels up to the heathen tomb, Cod. Dipl. Kmb. ii. 250, 13. (The same phrase often occurs in the charters in the descriptions of boundaries.) Se hæfde  
**rūm-gāl**; *adj.* *Rejoicing in ample space in which to move (applied to the dove when sent from the ark)*:—*Seó culufre wīde fleáh oþ ðæt heó*  
**sturtan** (*? vowel as in murnan?*); *steart To start, jump up*:—*Sturtende (sturtende (wk.)? v. examples from Middle English) se halta**

Figure 34 – Commentaries in the Printed Version, exemplified on “hæðen” (Bosworth and Toller, 1898, p. 502), “rūm-gāl” (Ibid., p. 804), and “sturtan” (Ibid., p. 930)

The three entries in Figure 34 all come from different parts of the dictionary (although all of them by Toller<sup>40</sup>) and appear at different levels of the microstructure, i.e. example, definition, and grammatical variants category respectively. What connects all these illustrations is the impracticability of any further markup being made to them, furthermore, they are in form longer and do not fulfil a particular lexicographic function (compare with intra-example glosses further into the chapter). As visible from the figure, a heterogeneous category is being dealt with, yet due to better readability and ease of editing, only a single element has to be assigned to the category. For the current format, such occurrences are contained in the element <comment> whose TEI Lex-0 counterpart would be <note>, therefore the correct markup is as follows<sup>41</sup>:

<sup>40</sup> Such occurrences (at least those marked by parentheses) are much rarer in the part of BT written by Bosworth. This may be due to the shorter nature of entries in Bosworth’s part.

<sup>41</sup> An argument could be made for the element <usg> instead of <note>, however the <usg> element is constrained to the “prototypical” labels “fig.,” “poet.,” etc...

```

<def>Rejoicing in ample space in which to move
  <comment>(applied to the dove when sent from the ark)</comment> </def>

<def>Rejoicing in ample space in which to move
  <note>(applied to the dove when sent from the ark)</note> </def>

```

Figure 36 – Comparison of Commentary Structure

The `<comment>` element does not play an important role in the user interface as it does not carry any graphic distinction, i.e. it retains the font of the parent element. In this case, as `<def>` carries the “basic” font, so would the additional information content of `<comment>`:

Figure 35 – Web Application Display of Commentaries, exemplified on “rúm-gál” (An Anglo-Saxon Dictionary Online, 2014, <https://bosworthtoller.com/26022>)

There is a subcategory of these comments/notes, that unlike the preceding illustrations, actually do contain information that may require additional markup. There has already been a case of this in the illustrative example that was intentionally left undescribed:

**I a. fig. to thrust one's self into the affairs of another, to exercise authority. v. in-, on-sting:—**Nā stinge nān mann on ðæt land, būton se hýred æt Xþes cyrcan, Chart. Th. 578, 6. Ic habbe ðæt geleornod, ðæt nān læwede man nāh mid rihte tō stingan hine on ānre cirican, nā an ān ðara ðinga ðe tō cyrcan belimþ. And for ðī wē forbeódaþ eallan læwedan mannum ænne hlāuordscipe ouer cyrcan, Cod. Dip. B. i. 137, 24. (Cf. Icel. þú hefir mjök stungizk til þessa máls thou hast meddled much with this case.) **II. to prick**

Figure 37 – Complex Commentary in the Printed Version, exemplified on “stingan” (Bosworth and Toller, 1898, p. 921)

This occurrence is particularly interesting because of several reasons; functionally the content is etymological and even the place in the structure corresponds to the etymological category,

however, there are two problems in interpreting this occurrence as purely etymological. Firstly, the meta marks “[ ]” (see Figure 18) are missing and there is a label of comparison “cf.” that may suggest incertitude of the etymological link between “stungizk” and “stingan”. Secondly, the form of the etymological information is atypical for a cognate structure (see the difference between cognate structure and reflex structures in chapter 8.3). Therefore, there are two plausible ways to mark up this given occurrence, either in the less precise but more economical way:

```
<comment>(Cf. Icel. Þú hefir mjök stungizk til þessa máls
  thou hast meddled much with this case.)</comment>
<note>(Cf. Icel. Þú hefir mjök stungizk til þessa máls
  thou hast meddled much with this case.)</note>
```

Figure 38 – Comparison of Simplified Markups for Complex Commentaries

or in a more precise, yet less economical way:

```
<etym>(Cf. <item><source>Icel.</source> Þú hefir mjök <cog>stungizk</cog> til þessa máls
  <trans lang="eng">thou hast meddled much with this case.</trans>.</item></etym>
<etym><lbl>Cf.</lbl> <lang>Icel.</lang> Þú hefir mjök
  <cit type="cognate"><form><orth>stungizk</orth></form></cit> til þessa máls
  <cit type="translation"><quote>thou hast meddled much with this case.</quote></cit></etym>
```

Figure 39 – Comparison of Complex Markups for Complex Commentaries

However, there are problems with both the possible solutions given. Starting with the comment interpretation, the lack of further markup goes against the underlying ideas of both formats. The current format strives for the best balance between faithfulness to the original and user-friendliness, which is not attained by the sole `<comment>` element. That is because both in the web application and in the original print, there is a graphic distinction between target languages (PDE and Latin) and source languages that is lost without the use of the element `<trans>`. For TEI Lex-0 the problem is simpler, the sole `<note>` element is not incorrect but would be deemed too general regarding the idea of the most precise markup in TEI Lex-0. Concerning the second solution, the problem for both formats is the same and that is the element `<etym>`. We are not given the meta marks and furthermore, we are encouraged to “compare”<sup>42</sup> the parenthesized text with the preceding examples in sense I a. Therefore, marking the text in parentheses as `<etym>` may be a reinterpretation or misunderstanding of the original which are to be avoided.

<sup>42</sup> Therefore `<see>` for current format or `<xr type=related>` for TEI Lex-0 may be more fitting. However, this markup would bring even more problems (not referencing an entry/sense, translations inside this element are deprecated, etc...) and is thus discarded.

One important thing to keep in mind is, that if an occurrence is not singular, it is important to come up with a standardized markup for all analogous occurrences. Therefore, before proceeding to the final version of the markup, see the analogous examples below:

**up-lendisc**; *adj.* *Uplandish, country* (as opposed to town), *rural, rustic*:—Uplendisc forensis (*forensis qui foras est*, Migne), Germ. 389, 41. Eft begann sum uplendisc mann egeslice hrýman tō ðám árleásu**m** burhwaru**m** . . . Ðá arn se ceorl geond ealle ða stræt hrýmende, Homl. Th. ii. 302, 4–8. Wē wyllaþ ðisue circul ámearkian, ðæt se uplendiscea preóst (cf. Chaucer's: Poure persoun dwelling uppon lond) wite his naman; mæg beón ðe glædre his heorte ðe hē sum þing hērof undergyte, **ge-wrixl**; *adj.* *Substitute: I. alternate. v. gewrixl(e); I a:—Gewrixlum sīþum alternis uicibus*, An. Ox. 7, 216: 8, 163. Stemnum (v. stefu a turn) gewrixlum, 3001. v. ge-wrixlic. **II. vicarious.**

Figure 40 – Further Complex Commentaries in the Printed Version, exemplified on “up-lendisc” (Bosworth and Toller, 1898, p. 1141) and “ge-wrixl” (Bosworth and Toller, 1921, p. 458)

The first illustration shows a very similar occurrence of “cf.” and (what would formally correspond to) etymology<sup>43</sup>, although structurally, we are mid-example, i.e. in a place where etymology does not belong. The second example is clearly a reference to an entry which is further described by its translational equivalent due to its homonymic nature<sup>44</sup> but once again in a place, where references do not structurally belong (compare “v. ge-wrixlic” at the end of sense I a – a structurally sound place for references, hence no parentheses).

Connecting the illustrations in Figure 40 to the preceding illustration, it is now clear that the structural soundness of the parenthetical text was by chance. Therefore, the negatives of the <etym> markup clearly outweigh the negatives of the <comment> markup, this becomes all the clearer when one of the main negatives of <comment> – the fact that the content is graphically indistinct from the surrounding text – can be diminished<sup>45</sup>. Getting back to the primary illustration (see Figure 37), the original version has the graphic distinction of the source language “Icel.” and PDE translation “thou hast meddled [...]” this can be easily retained through the elements <lang> and <trans> plus the element <cog> to keep in line with the rest of the current format. Therefore, the standardized way of marking up these occurrences could be described as: “element <comment> in which other child elements are nested, semantically corresponding to their structurally sound counterparts” as illustrated here:

<sup>43</sup> However, in this case of “reflex structure” etymology (see chapter 8.3).

<sup>44</sup> Compare ids 28844, 28845, and 28846.

<sup>45</sup> At present time, some element’s graphic distinctions are overruled by the <comment> font, for a better web application display the current XSLT document has to be tweaked (see chapter 9).



```
<comment>(Cf. <source>Icel.</source> Þú hefir mjök <cog>stungizk</cog> til þessa máls
<trans lang="eng">thou hast meddled much with this case</trans>.)</comment>
```

Figure 41 – Current XML Hybrid Markup of Complex Commentaries

and its TEI Lex-0 version:

```
<note>(<lbl>Cf.</lbl> <lang>Icel.</lang>
  Þú hefir mjök <cit type="cognate"><form><orth>stungizk</orth></form></cit>
  til þessa máls <cit type="translation"><quote>thou hast
  meddled much with this case</quote></cit>.)</note>
```

Figure 42 – TEI-Lex 0 Hybrid Markup of Complex Commentaries

However, due to practical reasons – the TEI Lex-0 markup being converted from the current format and the fact that the majority of BT has already been edited – also a bare element `<comment>` (which would convert to TEI-Lex 0 `<note>`) without any further nested elements may be at this point considered sufficient with the more rigid standardized markup being used once more salient issues are resolved. The second type of parenthesized texts in BT which should be held distinct to the commentaries is the intra-example gloss described in the following chapter.

## 8.2 Intra-example Glosses

Another marginal occurrence restricted mainly to the example part of the microstructure is the intra-example gloss which shares similarities with the definitional gloss as described above (see Figure 25). However, in this case, the `<gloss>` element contains not only additional descriptive information in parentheses but also specific contextual meanings and references to persons and objects. An illustration of such occurrences can be seen in the figure below:

**sele, es; m. A hall, house, dwelling** :—Cwom bytla (Guthlac) tō ðam beorge . . . wæs sele (his hermitage) nīwe, Exon. Th. 146, 24; Gū. 714. Sele sceal stondan, sylf ealdian, 343, 16; Gn. Ex. 158. Sele (Heorot, Hrothgar's hall) hlifade, heáh and horngéap, Beo. Th. 163; B. 81. Ðes sele, receda sēlest, 827; B. 411. Ðes windiga sele (hell), Cd. Th. 273,

Figure 43 – Intra-example Glosses in the Printed Version, exemplified on “sele” (Bosworth and Toller, 1898, p. 859)

The first gloss in the excerpt above is “(Guthlac)” and it comes after the word “bytla” (builder), therefore it is clear that the function of this gloss is referential as it reduces the semantic scope of “any builder” to one particular builder “Guthlac.” The second example “(his hermitage)” is a contextually specified type of dwelling, i.e. in connection to Guthlac the hermit, “sele” should not be translated/understood as a hall or a house but rather a hermitage.

The main difference between the referential and contextually specified meaning function is the positioning of the gloss. Whereas referential glosses follow a non-headword common noun or

pronoun, contextually specified meanings are most often found following the headword (in this case “sele”). Therefore, for simplicity’s sake, even the third and fourth gloss will be seen as contextually specified meaning, although one could (validly) argue that their functions are slightly different.<sup>46</sup>

Regarding the current XML format, the treatment of intra-example glosses has not been specified resulting in different practices by various editors. However, none of the current ways of markup can attain the underlying concept of truthfulness to the original as none of the child elements of <oe> carry a graphically distinct font<sup>47</sup>. Therefore, the only possibility to retain the graphical distinction was to divide the <oe> element and contain the gloss in a <trans> element such as is shown here with the web form below it:

```
<ex>
  <oe>Cwom bytla</oe>
  <trans lang="eng">(Guthlac)</trans>
  <oe>tó ðam beorge . . . wæs <cit>sele</cit></oe>
  <trans lang="eng">(his hermitage)</trans> <oe>níwe,</oe>
  <references><ref>Exon. Th. 146, 24</ref>; <ref>Gú. 714</ref>.</references>
</ex>
```

Cwom bytla  
(Guthlac)  
tó ðam beorge . . . wæs sele  
(his hermitage)  
níwe,

Exon. Th. 146, 24; Gú. 714.

Figure 44 – Current Format Markup of Intra-example Glosses and Its Web Application display

Nonetheless, neither the markup nor the web display are particularly suitable. An easy solution would be to tweak the XSLT document so that the element <trans> contained inside <oe> would carry the same font as if it is not contained. Yet, this would be a solution only for the end user and would undermine the effort of terminologically sound markup. Therefore, a new element for the current format is proposed to capture such occurrences, that is the element <gloss>. A question remains as to how to display this phenomenon in the web application – the idea of truthfulness to the original would dictate a singular graphic display for both referential glosses (as in bytla=Guthlac) and contextual meaning glosses (sele=hermitage). On the other hand, this brings issues for the end user, as such display may suggest that Guthlac is a type of builder<sup>48</sup> in

<sup>46</sup> In the case of the third gloss, it may be seen as a combination of the two functions. Regarding the fourth gloss, it possesses the specific function of a fixed expression – a kenning (windiga sele or wind-sele meaning hell)

<sup>47</sup> With the exception of <cit> which is reserved for the headword in the current format

<sup>48</sup> To make this even clearer, in some cases it may indicate that, for instance, Beowulf is a type of “he”



the same way as hermitage is a type of dwelling. Hence, the final proposition is to have two elements <gloss>, one bare, with a distinct font for contextual meanings, and one with an attribute (gloss func="ref") with a different font for references, optionally, both the elements may also possess the attribute "lang" analogically to the element <trans> as both these elements contain text in target languages. The XML version would then look like this:

```
<ex>
  <oe>Cwom bytla <gloss func="ref" lang="eng">(Guthlac)</gloss>
  tó ðam beorge . . . wæs <cit>sele</cit>
  <gloss lang="eng">(his hermitage)</gloss> níwe,</oe>
<references><ref>Exon. Th. 146, 24</ref>; <ref>Gú. 714</ref>.</references>
</ex>
```

Figure 45 – Updated Current Format Markup of Intra-Example Glosses

This type of markup, whilst being useful for the current format as it improves both the readability and the truthfulness to the original, draws largely from the TEI-Lex 0 principle of most precise markup, hence the TEI-Lex 0 format would look almost identical:

```
<cit type="example" xml:lang="ang">
  <quote>Cwom bytla <gloss type="ref" xml:lang="en">(Guthlac)</gloss>
  tó ðam beorge . . . wæs <ref type="oRef">sele</ref>
  <gloss xml:lang="en">(his hermitage)</gloss> níwe,</quote>
  <bibl>Exon. Th. 146, 24</bibl> <pc>;</pc>
  <bibl>Gú. 714</bibl> <pc>.</pc>
</cit>
```

Figure 46 – TEI-Lex 0 Markup of Intra-Example Glosses

This means that once the new element is implemented to the current schema and utilized in the current XML documents, the subsequent TEI-Lex 0 conversion will not pose any problems. However, for the earliest stages of TEI-Lex 0 conversion, treating <trans> graphically in the same way whether it stands on its own or is nested in <oe> should suffice. It should not pose any problems to the validity of the converted TEI-Lex 0 as <cit type="translation"> can be a child of <cit type="example">. Nonetheless, the suggested updated markup for the current format should be superior and should allow for a TEI-Lex 0 conversion more in line with its idea of the most precise markup, and thus such a markup is suggested to be included in the next round of editing.

The second phenomenon studied in this chapter is the intra-example variant, which similarly to the preceding case, illustrates an occurrence prototypically found in the definitional part (<form type="variant"> in TEI-Lex 0 and <var> in the current format). At first glance, the intra-example variant is distinct from the intra-example gloss due to the employment of solely the

source language and a reference to the source text where the given variant arises. A prototypical example of the intra-example variant would therefore look like this:

**the sentence and (a) following it immediately:—Ic sylf̄ (seolf, Lind.:  
solfa, Rush.) hit eom ipse ego sum, Lk. Skt. 24, 39. Heó sylf hié þeówen**

Figure 47 – Intra-example Variant in the Printed Version, exemplified on “self” (Bosworth and Toller, 1898, p. 860)

This illustration comes from the entry id: 27390 under the lemma “self”. Hence, even the example’s base form “sylf” is an orthographic variant of the lemma – the relationship between the lexeme form in the example and the lemma is signaled by the use of element <cit> in the current format and <ref type=oRef> in TEI-Lex 0. The parenthesized text is then understood as further variants of the same sentence in different sources, i.e. “Ic seolf hit eom” should exist in Lindisfarne gospels and “Ic solfa hit eom” in Rushmore gospels. This structure could be marked up followingly in the current format:

```
<ex>
  <oe>Ic <cit>sylf</cit>
    <comment>(<cit>seolf</cit>, <ref>Lind.</ref>: <cit>solfa</cit>, <ref>Rush</ref>.)</comment>
    hit eom</oe>
  <trans lang="lat">ipse ego sum,</trans>
  <references><ref>Lk. Skt. 24, 39</ref>.</references>
</ex>
```

Figure 48 – Current Format Markup of Intra-example Variants

Unfortunately, such a markup brings more issues than solutions; for one, the element <comment> has already been restricted to annotations requiring no further markup and to annotations beginning with a referential label (v. and cf.), moreover, there is no clear relationship between the parenthesized variant and its source, i.e. there is no parent element nesting these two elements. Both of these problems can be simply solved, firstly by the employment of a new element <gloss type="variant"> instead of <comment> and secondly by treating the variant as an example of its own, such a markup would look like this:

```
<ex>
  <oe>Ic <cit>sylf</cit>
    <gloss type="variant">(<ex><oe><cit>seolf</cit></oe>, <ref>Lind.</ref></ex>:
      <ex><oe><cit>solfa</cit></oe>, <ref>Rush</ref>.</ex>)</gloss>
    hit eom</oe>
  <trans lang="lat">ipse ego sum,</trans>
  <references><ref>Lk. Skt. 24, 39</ref>.</references>
</ex>
```

Figure 49 – Updated Current Format Markup of Intra-example Variants

And its TEI-Lex 0 counterpart:

```
<cit type="example" xml:lang="ang">
  <quote>Ic <ref type="oRef">sylyf</ref>
    <gloss type="variant">(<cit type="example"><quote><ref type="oRef">seolf</ref></quote>
      <pc>,</pc> <bibl>Lind.</bibl></cit>
      <pc>:</pc>
      <cit type="example"><quote><ref type="oRef">solfa</ref></quote>
        <pc>,</pc> <bibl>Rush</bibl></cit> <pc>.</pc>)</gloss>
    hit eom</quote>
  <cit type="translation"><quote>ipse ego sum,</quote></cit>
  <bibl>Ik. Skt. 24, 39</bibl><pc>.</pc>
</cit>
```

Figure 50 – TEI-Lex 0 Markup of Intra-example Variants

This type of markup ensures a simple 1:1 conversion from the current schema to the TEI-Lex 0 schema and with a user-oriented HTML transformation, it would also improve the readability of the digitized dictionary as many of the advantages of further markup in the parenthesized text would be retained (e.g. graphical distinction of <cit> or hyperlink function of <ref>).

The last two chapters were preoccupied with the parenthesized texts in BT that follow a regular structure, yet currently lack a standardized markup. This issue has been solved by adding new elements to the current format that are based on TEI-Lex 0. A summarization table of the elements for parenthesized texts, their usage, and their TEI-Lex 0 counterpart can be found below:

Current format element	TEI-Lex 0 element	Usage	Example
<comment>	<note>	Any unstructured parenthesized text that does not require further markup.	Ðæt wif <comment> (wif though neuter is represented by a fem. pron.) </comment>
<comment> + further markup	<note> + further markup	Parenthesized text starting with a referential label and requiring further markup	Sincfæt <comment> (<lbl>cf.</lbl> fæted wáge, <ref>4553</ref>; <ref>B. 2282</ref>) </comment>
<gloss>	<gloss>	Parenthesized text that functions as a context-specific translation	Sinnehte <gloss lang="eng"> (hell) </gloss>
<gloss type="ref">	<gloss type="ref">	Parenthesized text that functions as a reference to a particular person or object	Him <gloss type="ref" lang="eng"> (Beowulf) </gloss>
<gloss type="var">	<gloss type="var">	Parenthesized text that functions as an orthographic variant	syleþ <gloss type="var"> (<ex><oe> <cit>selleþ</cit></oe>, <ref>Rush.</ref></ex>) </gloss>

Figure 51 – Parenthesized Structures Typology

These elements account for the absolute majority of all parenthesized texts in the BT and having a standardized markup with a user-centered HTML transformation should further improve the readability and truthfulness-to-original of the digitized version. The proposed markup is however a novelty and has not been used in the currently edited entries, therefore at the beginning of the TEI-Lex 0 conversion, such occurrences will be in TEI-Lex 0's terms too simplified<sup>49</sup>. The proposed updated current format would solve these issues but would be considerably more time-consuming. Having dealt with the parentheses, it is now time to turn

<sup>49</sup> The described structures may currently be marked up by a bare <comment> or not marked at all, resulting in the "too simplified" markup in TEI-Lex 0 which would be <note> or respectively no markup. This TEI-Lex 0 markup would however still be valid against its schema.

non-parenthesized structures that would benefit from a new markup in the current format that is, again, based on the TEI-Lex 0 markup.

### 8.3. Cognates, Reflexes, and Etymons

The first structure that would benefit from a further markup distinction is the etymological category. Currently, there is no differentiation regarding the three etymological forms of BT; these forms can be found in the example below with “reflex structure” underlined in green, “cognate structure” in blue, and “etymon structure” in red:

[Hi harm hadde, hii wende þat hii siker were, Laym. 9401 (2nd MS.). Dead is þe king, & siker þu miht hider comen, 15092. Wā wes Brutten þere, þenne heo wenden beon sikere, 29289. Be þu sikerr þatt he shall þe 3ifenn eche blisse, Orm. 4844. Beoð ancren wise, þet habbed wel bituned ham a3ein þe helle leun, uorte beon þe sikerure, A. R. 164. 12. Ne migten he siker ben, for magnie of ðo woren ouertaken, Gen. and Ex. 876. þat ich mowe a siker bold arere, R. Glouc. 116, 1. Syker þou be Engeland ys nou þyn, 359, 9. Hit is sikerest in þi heued (*safest to sprinkle water on the head at baptism*), Shoreham. þai salle be þare syker and certayne To have endeless joy, Pr. C. 8559. A man hath most honour To deyen . . . whan he is siker of his goode name, Chauc. Kn. T. 2191. Her none sikerer þan other, Piers P. 12, 162 note. O. Frs. sikur (-er) *free from guilt; sure, trustworthy*: O. Sax. (sundiono) sikur (-or): O. H. Ger. sihhur securus, immuns, liber, tutus. From Latin securus.]

Figure 52 – Various Etymology Structures in the Printed Version, exemplified on “sikor” (Bosworth and Toller, 1898, p. 870)

These forms differ both structurally and functionally. The reflex structure consists of an example sentence (containing the reflex) in the initial position with a reference to a specific part of the book in the final position and its function is to give etymological information from later-stage English (mostly ME). The cognate structure begins with the abbreviation of a language followed by a cognate from this language with an optional translation equivalent at the end of the structure; functionally, this structure conveys etymological information regarding (mostly Germanic) contemporary languages of OE. The “etymon structure” is the rarest of the three and is formed by the source preposition “From” followed by the source language and the etymon<sup>50</sup>, the function of this structure is to give the source lexeme from which the OE form descends, i.e. its function is the opposite of the reflex.

<sup>50</sup> Etymon, in the strictest sense, is the ultimate source, i.e. the PIE form of the lexeme. For the purposes of this paper, any structure beginning with the source preposition “From” is understood as the etymon structure (this consists mainly of forms “From Latin”, “From Hebrew”, and “From Greek”).

Notwithstanding the functional and structural differences, the current format offers a singular way of markup. Taking one of each of the structures in Figure 52, the current XML version would look followingly<sup>51</sup>:

```
<etym>[<item>Hi harm hadde, hii wende þat hii <cog>siker</cog> were,
  <source>Laym. 9401 (2nd MS.)</source>.</item>

  <item><source>O. Frs.</source> <cog>sikur</cog> <cog full="siker">(-er)</cog>
    <equiv lang="eng">free</equiv> from guilt; <equiv lang="eng">sure</equiv>,
    <equiv lang="eng">trustworthy</equiv></item>

  <item>From <source>Latin</source> <cog>securus</cog>.</item>]</etym>
```

Figure 53 – Current Format Markup of Various Etymology Structures

The solution to this problem is quite simple, either we can define new child elements of <item> to be <reflex> and <etymon>, so that <cog>, <reflex>, and <etymon> are used in their respective structures or allow for attributes on the element <item>, so that we have three distinct forms of the element, e.g. bare <item> for cognate structure as it is the most common, <item func="reflex"> for reflex structure and <item func="etymon"> for etymon structure. Basing the decision on the easiest possible conversion to TEI-Lex 0, the adding of attribute to <item> has been elected. For the reflex structure, as we are dealing with a complete sentence with a reference to a particular book, <source> element was substituted by <ref> and the whole sentence is embedded in <quote> element analogically to citations in the example category. The new way of markup in the current schema would therefore look like this:

```
<etym>[<item function="reflex"><quote>Hi harm hadde, hii wende þat hii <cog>siker</cog> were,</quote>
  <ref>Laym. 9401 (2nd MS.)</ref>.</item>

  <item><source>O. Frs.</source> <cog>sikur</cog> <cog full="siker">(-er)</cog>
    <equiv lang="eng">free from guilt</equiv>; <equiv lang="eng">sure</equiv>,
    <equiv lang="eng">trustworthy</equiv></item>

  <item func="etymon">From <source>Latin</source> <cog>securus</cog>.</item>]</etym>
```

Figure 54 – Updated Current Format Markup of Various Etymology Structures

However, considering the conversion TEI-Lex 0 we still run into a problem caused by the limitations of the standard. The counterpart to current format <item func> in TEI-Lex 0 is <cit type>, yet the only defined attribute values to <cit type> regarding etymology are “etymon”, “cognate”, and “cognateSet”. Therefore, to signalize reflex structures in TEI Lex-0 the proposed markup is <cit type="cognate" subtype="reflex">, as this is not a predefined element, it has to be mentioned in the TEI header. The corresponding TEI Lex-0 format is the following:

<sup>51</sup> The element <cog> currently does not allow any attributes, see “further changes” for the element <cog full> visible in the cognate structure (O. Frs.).

```

<etym>[<cit type="cognate" subtype="reflex"><quote>Hi harm hadde, hii wende þat hii
  <hi>siker</hi> were,</quote> <bibl>Laym. 9401 (2nd MS.)</bibl><pc>.</pc></cit>

  <cit type="cognate"><lang>O. Frs.</lang> <form><orth>sikur</orth></form>
    <form><orth expand="siker">(-er)</orth></form>
    <cit type="translationEquivalent"><orth>free from guilt</orth></cit>
    <pc>;</pc><cit type="translationEquivalent"><orth>sure</orth></cit>
    <cit type="translationEquivalent"><orth>trustworthy</orth></cit></cit>

  <cit type="etymon"><lbl>From</lbl> <lang>Latin</lang>
    <form><orth>securus</orth></form><pc>.</pc></cit>]</etym>

```

Figure 55 – TEI-Lex 0 Markup of Various Etymology Structures

It has to be noted that both the updated current format and the TEI-Lex 0 are simplified in regard to multi-word definitions. For instance, in the phrase “free from guilt” the most precise markup would differentiate between the translation equivalent “free” and the explanation “from guilt” (as distinguished in the original through the use of italics and base font: “*free* from guilt”). The simplified markup was chosen as the current format would not allow for a TEI-Lex 0 conversion without major changes to the current hierarchy which is sufficient for the needs of the web application. All in all, the updated markup would facilitate a more rigid version of BT which at the same time would improve the readability for the users of the web application<sup>52</sup>.

With the typology of etymological forms out of the way, all the various structures of BT have been depleted and sufficiently described both in TEI-Lex 0 and the updated current format. The focus will now turn to a list of further changes that will not require such a comprehensive description as the previous structures.

---

<sup>52</sup> For example, <cog> in <quote> would be rendered differently than when embedded in <item> as it is harder to spot in a sentence (as in reflex structures) than on its own (as in cognate and reflex structures).



## 9. Further Changes

This chapter will describe some lesser changes to the current format or the way the current format is transformed into the HTML webpage. These changes should not in any way hinder the TEI-Lex 0 conversion nor make it simpler as the focus of this chapter is to give the user of the web application the most user-friendly and the most true-to-original variant of the BT.

First of these changes has already been hinted at in the preceding chapters and that is the graphical distinction of elements nested in another element (most often `<comment>`). The three elements needing an updated transformation are `<trans>`, `<references>`, and `<see>`. Starting with the element `<trans>`, it is the only font-carrying element whose distinction is lost when nested in another element. An example<sup>53</sup> of this compared against other font-carrying elements can be seen in the figure below with the XML markup at the top and the current HTML transformation at the bottom:

```
<ex> There is a <equiv lang="eng">translational equivalent</equiv>, <cit>form of the headword</cit>,
<cog>cognate</cog> and a <trans lang="eng">translation</trans>
<comment>(this is the <equiv lang="eng">translational equivalent</equiv>,
this is <cit>form of the headword</cit>,
this a <cog>cognate</cog>,
and finally a <trans lang="eng">translation</trans>)</comment></ex>
```

There is a translational equivalent, form of the headword, *cognate* and a translation (this is the translational equivalent, this is form of the headword, this a *cognate*, and finally a translation)

Figure 56 – Font-carrying Elements and Their Web Application Display

The solution to this problem is simple as the only thing that has to be done is updating the transformation of `<trans>` in line with the other elements shown in the figure. Whilst `<trans>` can be said to be lacking in graphical distinction when nested in another element, the opposite can be said about the elements `<see>` and `<etym>`. The difference between these elements is that `<trans>` is a font-carrying element whilst `<see>` and `<references>` are category elements, i.e. they are transformed into list-like structures in the web application which improves the readability of the dictionary when the categories occupy their prototypical place but hinder it anywhere else as visible from the figure:

---

<sup>53</sup> This is a theoretical example, the text shown is not part of the actual BT.



```

<examples><ex><oe>
  This is an example sentence <comment>(<see>v. <a href="12345">death-sentence</a></see>
    as used in <references><ref>YX 24, 42</ref>.</references>)</comment></oe>
  <trans lang="lat">Hoc est exemplum sententia</trans>
</ex></examples>
<see>v. <a href="12345">death-sentence</a></see>

```

The figure consists of two rectangular panels. The top panel has a green border and is titled 'Transformation when nested' in the top right corner. It displays the text 'This is an example sentence (' followed by a blue link icon and 'Linked entries'. Below this, 'v. death-sentence' and 'as used in' are shown as blue hyperlinks. The text 'YX 24, 42.' is also shown as a blue hyperlink. The panel ends with a closing parenthesis ')'. The bottom panel has a blue border and is titled 'Transformation when non-nested' in the bottom right corner. It displays the text 'Hoc est exemplum sententia' followed by 'XY. 12, 34'. Below this, there is a blue link icon and 'Linked entries', followed by 'v. death-sentence' as a blue hyperlink. The text 'XY. 12, 34' is not a hyperlink.

Figure 57 – Category-defining Elements and Their Web Application Display

There are two possible solutions to this problematic transformation; either the XSLT document has to be updated so that the elements `<see>` and `<references>` carry no graphical distinction when embedded in another element or to standardize markup without these elements (when nested) and let the hyperlink function be carried by `<a href>` or `<ref>` respectively. As of now, if `<a href>` and `<ref>` are not nested in their category denoting parent elements, they do not show as hyperlinks in the web application. Possibly, both of these solutions should be reflected in the updated XSLT document as the editing practice for such structures has not been standardized and both ways of markup, i.e. “`<comment>(<see> v. <a href=“123”>reference entry</a> <references><ref>source</ref></reference>)</comment>`” and “`<comment>(v.<a href=“123”>reference entry</a> <ref>source</ref>)</comment>`” may have been used by various editors. Both changes would lead to the same web application interface and improve its user-friendliness.

The next proposed change concerns the sense markings in the current format and the possibility of hyperlinks to specific senses rather than entries. The numbering of senses is currently dual, i.e. once signaled through the attribute value of `<sense num=“X”>` and then by the element `<snum>` containing the actual text of the dictionary, i.e. `<snum>X</snum>`. The current

transformation takes the attribute value as the sense marker and omits the text contained in `<snum>`. An illustration can be seen here:

```
<sense num="I">
  <snum>123456789</snum>
  <def>this is an example</def>
</sense>
```

I. this is an example

Figure 58 – Sense-numbering Elements and Their Web Application Display

The first step of facilitating sense reference would be to swap the functions of the attribute value and the text contained in `<snum>`, taking the preceding figure as example, it would be the number “123456789” that the end user would see and not the “I”. The second step would be to add the entry id (for example `<entry id=“98765”>`) in which the `<sense num=“I”>` is nested into the attribute value, distinguishing between entry number and sense number by any character, for example underscore. This process can be automated and would leave us with a unique ID for every sense (in our example `<sense num=“98765_I”>`). Such markup would then facilitate better reference in cases such as this one, where a specific sense is in question:

**ge-wyrdelic.** *Add: I. fortuitous:—Gewyrdelicum gelimpe fortuitu casu, An. Ox. 3792. Þā gewyrdelican āwendennessa fortuitas permutaciones, 190. II. of narrative, recording events, historical. v. gewyrd; III:—Fram gewyrdelicere race ab istorica relatione, An. Ox. 3028. Hyra ôðer āwrāt þās gewyrdelican race, Hml. S. 6, 366.*

Figure 59 – Reference to a Specific Sense in the Printed Version, exemplified on “ge-wyrdelic” (Bosworth and Toller, 1921, p. 461)

An easier navigation to the specific passage would increase the quality of life for the users of the web application, the only problem is, that while the sense numbering can be automated, the reference to that particular sense number would have to be done manually. This task can be made easier by searching for particular characters inside the `<see>` element, e.g. Roman numerals, single alphabetical characters and single characters from the Greek alphabet, but it still would be one of the more time-consuming tasks suggested in this chapter.

The last suggested change to the current format is the addition of the optional attribute “full” to the element `<cog>`. This change would be in line with the current format as all other possibly abbreviated lexeme forms possess this attribute (`<infl full>` and `<var full>`). While cognates are not shortened as often as the other forms mentioned, in entries where they do occur the

suggested <cog full> markup would bring the same advantages to the reader as the two aforementioned forms. An example of such an entry can be found below:

[O. H. Ger. *hagazissa* (-ussa) *furia*; *hāzus* (-is) *strihia*, *erynnis*: Ger. *hexe*.] Cf. *heah-rūn*.

Figure 60 – Orthographic Variants of Cognates in the Printed Version, exemplified on “hægtesse” (Ibid., p. 495)

Where the underlined cognate forms would be marked up as <cog full=“hagazussa”>(-ussa)</cog> and <cog full=“hāzis”>(-is)</cog>. The only negative is that this markup would have to be done manually, on the other hand, as mentioned, there are not as many abbreviated forms of cognates and all of them can be found quite simply by searching for the string of characters “(-” under the etymology element.

To recapitulate, this chapter was focused on changes to the current format and its transformation to HTML web application which would improve the user-interface while not being overly difficult to set up. The first suggested change was to keep the distinct font of element <trans> even if it is nested in other elements. Then we moved to the elements <see> and <references> where it was suggested that they should lose some of their graphic properties when embedded in another element. The next change concerned the possibility of reference to unique senses of an entry rather than the whole entry – this task was twofold, the automated part would be the conversion of non-unique sense ids to unique ones, the manual version would then consist of marking up the specific references whose target is the sense rather than the entry. And the last possible update was the inclusion of attribute “full” in element <cog> in order to keep the dictionary consistent with its already established forms <infl full> and <var full>.

## 10. Summary

The aim of this diploma thesis was to capture the almost 200-year history of *An Anglo-Saxon Dictionary* by Bosworth and Toller. During its conception, BT was regarded as a failure of a dictionary that would soon be superseded by an overall better dictionary. However, it has now been more than a century since the last addition to BT by Toller, and yet, there has not been a finalized project that would take BT's place. The historical impact of BT and the fact that no currently finalised dictionary comes close in comprehensiveness were the two main reasons that led to the digitization of the dictionary. This task was first taken up Sean Crist who began the digitization project and then transferred under the Faculty of Arts of Charles University where it has been continued under the leadership of Ondřej Tichý. During this period, many improvements have been made such as the transfer of data to the XML or the creation of the web application.

The next section of the paper referred about the preservation of digital objects. In the opening section, it was shown that preservation in respect to digital objects entails three other important notions – the data have to be stored in an accessible place, the data itself must be accessible, i.e. readable and the digitized object is enhanced in ways that make it useful for the modern user. The issue regarding the readability of the data was further discussed in chapters on the XML standards used to mark up data in the field of humanities. The most frequently used standard in humanities is the TEI which, if used for BT, would greatly help in its preservation. However, the lenient structure of TEI was found to not be the best choice for BT as a better variant guaranteeing not only better preservation but also better inter-operability between BT and other dictionaries. This version is called TEI-Lex 0, it is a substandard of TEI devised only with dictionaries in mind and offering a more rigid XML structure specialized to include majority, if not all, lexicographic structures and hence was chosen as the standard of choice for BT.

The next chapters were devoted to the analysis of BT. First, BT was compared against its supposed successor in DOE to find out whether the older BT still has something to offer that the modern DOE does not include. Starting with the comparison of the number of entries in both dictionaries, the apparent superiority of BT was quickly dissipated as both dictionaries were found to use different methods in lemmatization, i.e. despite BT having more entries than DOE, the OE lexis is represented in similar depth in both the dictionaries as the BT's additional entries are listed in DOE as orthographic variants under different headwords. In terms of common nouns, DOE was found to be the more comprehensive resource, however, in terms of proper nouns and affixes BT is the better source. Hence, already the first comparison showed that DOE is not simply a substitute for BT as either of the dictionaries has advantages of its

own. As for entry content comparison, DOE is the more comprehensive of the two in majority of the categories studied, yet, in some of the categories it is lacking in comparison to BT. One of these categories is the orthographic variant and inflected form category, where BT lists the grammatical environments in which the inflected forms can be found whilst DOE lacks this information in many of the entries. Another, even more important feature present in BT but missing in DOE is the translation category. Whereas BT gives a PDE or Latin translation for the majority of its Old English examples, DOE gives translations only for the restricted number of bilingual Latin-Old English texts with no PDE translations whatsoever. In this regard, BT was found to be the more useful resource for the general public or researchers with limited proficiency in Old English. All in all, it was postulated that digitization and preservation of BT is a task worth pursuing not only because it will not be simply substituted by DOE due to the reasons above but also because of its historical importance, its impact on the field of Old English lexicography and for the practical reason that loss of a dictionary such as BT, listed as a source in many other studies would render the citations impossible to check and hence making the whole field of Old English lose its integrity.

The following chapter discussed the premise upon which the digitization process of BT was built. The premise is to keep the digitized version as true-to-original whilst making it a user-friendly tool applicable for researchers of present and future. The balance is attained by making no changes to the actual text of BT whilst making use of the technological advancements and additional features offered by the digital medium that were during the conception of paper BT unthinkable. These features include toggle functions that can transform the original orthography using acutes into the currently preferred orthography employing macrons instead or into a version with no diacritics whatsoever. Another important addition is the employment of hyperlinks that makes navigation between various entries in BT or even the navigation between BT and another primary source much easier and user-friendly. The last improvement mentioned in the paper concerned the graphical distinctions between the various structures found in BT. Whilst paper medium permitted only a very restricted set of fonts and graphically distinct structures, the graphical possibilities for digital media are infinite, therefore, each structure, each occurrence with a specific function can be assigned a unique graphical display which in turn makes the web application much easier to navigate through.

The next section was devoted to the in-depth analysis of all the prototypical structures found in the original BT. The hierarchy of structures was presented and so were the graphical means by which these structures are originally distinguished. Then it was shown how this hierarchy translates to the currently used XML markup validated against the custom schema designed for

the specific needs of BT. The other thing discussed was the way in which the XML markup is transformed via an XSLT document to the HTML-based web application and a clear line was drawn between specific elements and specific graphical distinctions. Lastly, through synthesis of the notions mentioned in the chapter on preservation and the description of the current XML format, a new format utilizing the standardized markup of TEI-Lex 0 was devised. Every prototypical structure of BT was tagged in a TEI-Lex 0-conformant way and compared to the current tagging. Similarities and differences between the two types of markup were described and when one-to-one conversion was found to be impossible, solutions to the problems were proposed. The result of this chapter was a conversion table that would later be used as the basis for the XSLT document that would facilitate a simple one-to-one conversion between the current format and the more preservable and inter-operable TEI-Lex 0 format.

The following chapters presented some more complex and less common structures of BT. As these structures lack unified markup in the current format, TEI-Lex 0 conversion would be impossible as all the possible combinations of current elements used to tag such marginal structures would have to be converted to a single combination of TEI-Lex 0 elements. The marginal structures described were commentaries, intra-example glosses and variants, and various etymological structures. Commentaries were further divided into bare comments and complex comments requiring further markup (typically starting with a referential label “v.” or “cf.”). Glosses were assigned a specific type based on their function, referential for references to proper nouns, context translational for specific meanings of the headword based on the context of the specific example, and varietal for further orthographic variants of the headword. Lastly, the etymological structures were distinguished into cognate structures for language contemporaries of OE, reflex structures for later-stage English forms of the headword, and etymon structures for the source language of the given OE headword.

The last chapter served as a list of minor changes to the current XML format and XSLT document that would improve the readability of the web application whilst not being as time-consuming to set up as the new unified markup of complex structures described in the preceding chapters. These changes consisted of different graphic distinctions of various elements when nested inside another element and a proposition of a new sense marking that would facilitate both sense references and entry references as opposed to the current markup that permits only entry references.

Whilst there is still so much more that can be done to enhance the digitized version of the dictionary, e.g. merging the main volume and supplementary entries while abiding Toller’s editing information, describing even more marginal structures, or coming up with an automated

script that would eliminate the need of manual tagging, I believe that the purpose, with which this paper was conceived, was fulfilled and we are one step closer to a more preservable, interoperable and user-friendly version of Bosworth and Toller's *Anglo-Saxon Dictionary*.

## 11. Shrnutí v Českém Jazyce

Cílem této diplomové práce bylo zachytit téměř 200letou historii *Anglosaského slovníku* (BT), jehož autory byli Joseph Bosworth a Thomas Northcote Toller. V době svého vzniku byl BT považován za neúspěšný slovník, který je odsouzen k tomu, aby byl co nejdříve nahrazen staroanglickým slovníkem po všech stránkách lepším. Je tomu již však více než sto let, co Toller publikoval svůj *Dodatek k Anglosaskému slovníku*, a přesto dosud neexistuje dokončený projekt jež by BT předčil a nahradil. Byl to právě historický dopad BT a skutečnost, že žádný v současnosti dokončený slovník staré angličtiny se mu nepřibližuje svou komplexností, které vedly k rozhodnutí o digitalizaci slovníku. Tohoto úkolu se nejprve ujal Sean Crist, který projekt digitalizace zahájil v roce 2001, v roce 2006 následovně projekt přešel pod Filozofickou fakultu Univerzity Karlovy, kde práce na digitilazici pod vedením Ondřeje Tichého stále pokračuje. Během tohoto období došlo k mnoha vylepšením, jako je převod dat do XML nebo vytvoření webové aplikace.

Práce se dále týkala uchovávání digitálních objektů. V úvodní části bylo ukázáno, že uchovávání ve vztahu k digitálním objektům zahrnuje tři důležité úkony – data musí být uložena na přístupném místě, data samotná musí být přístupná, tj. čitelná, a digitalizovaný objekt by měl modernizován tak, aby byl užitečný pro novodobého uživatele. Otázkou týkající se čitelnosti dat jsme se dále zabývali v kapitolách o standardech XML používaných pro označování dat v oblasti humanitních věd. Nejčastěji používaným XML standardem v humanitních vědách je TEI, který by v případě využití v BT výrazně pomohl při jeho uchovávání. Ukázalo se však, že flexibilní struktura TEI není pro BT tou nejlepší volbou, neboť existuje lepší varianta zaručující nejen jednodušší uchování, ale také lepší interoperabilitu mezi BT a jinými slovníky. Tato varianta se nazývá TEI-Lex 0, jež je verzí TEI navrženou pouze s ohledem na slovníky, která nabízí rigidnější XML strukturu specializovanou čistě na lexikografické struktury. TEI-Lex 0 byl tedy následně vybrán jako nejvhodnější standard, do kterého by se nynější data konvertovala.

Další kapitoly byly věnovány analýze BT. Nejprve byl BT porovnán se svým předpokládaným nástupcem *Slovníkem staré angličtiny* (DOE), aby se zjistilo, zda má starší BT stále co nabídnout ve srovnání s modernějším DOE. Na začátku to vypadalo, že je BT mnohem obsáhlejší než DOE, protože jeho počet hesel byl mnohem vyšší. Tato představa se však rychle rozplynula, neboť se ukázalo, že oba slovníky používají při lemmatizaci odlišné metody, tj. přestože BT má více hesel než DOE, lexikum OE je v obou slovnících zastoupeno v podobné šířce, neboť hesla, která jsou v BT „navíc“, jsou v DOE uvedena jako ortografické varianty jiného hesla. Co se týče hesel popisujících obecná podstatná jména, DOE byl shledán obsáhlejší



zdrojem, avšak pokud jde o hesla popisující vlastní jména a afixy, je lepším zdrojem BT. Již první srovnání tedy ukázalo, že DOE není pouhou substitucí BT, protože oba slovníky obsahují kategorie, ve kterých jsou obsáhlejší než jejich protějšek.

Při porovnání obsahu hesel se ukázalo, že ve většině kategorií nabízí DOE vyšší komplexitu, ale jsou i takové kategorie, kde je komplexnějším slovníkem BT. Jednou z těchto kategorií je kategorie ortografických variant a skloňovaných tvarů, kde BT uvádí gramatická prostředí, v nichž se skloňované tvary vyskytují, zatímco DOE tuto informaci v mnoha heslech postrádá. Dalším, ještě důležitějším prvkem, který je v BT přítomen, ale v DOE chybí, je překladová kategorie. Zatímco BT uvádí u většiny svých staroanglických příkladů překlad do soudobé angličtiny nebo latiny, DOE uvádí překlady pouze u omezeného počtu dvojjazyčných latinsko-staroanglických textů a překlady do soudobé angličtiny chybí. V tomto ohledu byl BT shledán užitečnějším zdrojem pro širokou veřejnost nebo badatele s omezenou znalostí staré angličtiny. Výsledkem této části práce bylo, že digitalizace a uchování BT je úkol, který stojí za to realizovat a to nejen proto, že, jak bylo výše ukázáno, DOE nedokáže BT po všech stránkách plně nahradit, ale také kvůli jeho historickému významu, jeho vlivu na obor staroanglické lexikografie a z praktického důvodu, že ztráta slovníku, jako je BT, uváděného jako zdroj v mnoha dalších studiích, by znemožnila kontrolu citací, a tím by integrita celého oboru staré angličtiny byla narušena.

Následující kapitola pojednává o předpokladech, na nichž byl proces digitalizace BT postaven. Základem je, aby digitalizovaná verze zůstala co nejvěrnější originálu a zároveň se stala uživatelsky přívětivým nástrojem použitelným pro badatele i širokou veřejnost. Této rovnováhy je dosaženo tím, že zatímco vlastní text BT se nemění, využívá se jiných funkcí, které digitální médium nabízí a které byly v době vzniku papírové verze BT nemyslitelné. Mezi tyto funkce patří „přepínací“ funkce, která dokáže původní ortografii používající akuty změnit na v současnosti preferovanou verzi využívající makrony nebo na zjednodušenou verzi bez diakritiky. Dalším důležitým doplňkem je použití hypertextových odkazů, které výrazně usnadňují a uživatelsky zpříjemňují navigaci mezi různými hesly v BT nebo dokonce mezi BT a jiným primárním zdrojem zkoumající stejnou problematiku. Poslední vylepšení zmíněné v práci se týkalo grafického rozlišení různých struktur, které se v BT nacházejí. Zatímco papírové médium umožňovalo pouze velmi omezenou sadu fontů a grafického odlišení rozdílných struktur, grafické možnosti digitálního média jsou nekonečné, a proto lze každé strukturu, každému výskytu se specifickou funkcí přiřadit jedinečné grafické zobrazení, které následně značně usnadňuje orientaci ve webové aplikaci.

Další část byla věnována hlubší analýze všech prototypických struktur v původním BT. Byla představena hierarchie struktur a také grafické prostředky, kterými jsou tyto struktury rozlišeny. Poté bylo ukázáno, jak se tato hierarchie převádí do aktuálně používaného značení XML validovaného podle schématu navrženého pro specifické potřeby BT. Dále se hovořilo o způsobu, jakým se značení XML transformuje prostřednictvím dokumentu XSLT do webové aplikace založené na HTML, a byla jasně vymezena hranice mezi konkrétními XML elementy a konkrétním grafickým rozlišením. V rámci této kapitoly byl také navržen nový formát, který používá značení dle standardu TEI-Lex 0. Každá prototypická struktura BT byla označena způsobem odpovídajícím TEI-Lex 0 a porovnána se současným značením. Byly popsány podobnosti a rozdíly mezi oběma typy značení, a pokud se ukázalo, že převod jedna ku jedné není možný, byla navržena řešení problémů. Výsledkem této kapitoly byla konverzní tabulka, která bude později sloužit jako základ pro XSLT dokument, který by umožnil jednoduchý převod jedna ku jedné mezi současným formátem a formátem TEI-Lex 0.

V následujících kapitolách byly představeny některé složitější a méně obvyklé struktury BT. Protože tyto struktury nemají v současném formátu jednotné značení, konverze do TEI-Lex 0 by nebyla možná, protože všechny možné kombinace současných elementů používaných k označování takových marginálních struktur by musely být převedeny na jedinou kombinaci elementů TEI-Lex 0. Popsané netypické struktury byly komentáře, glosy a varianty uvnitř příkladu a různé etymologické struktury. Komentáře byly dále rozděleny na jednoduché komentáře a komplexní komentáře vyžadující další značení (obvykle začínající odkazovou značkou "v." nebo "cf."). Glosám byl přiřazen specifický atribut na základě jejich funkce, referenční pro odkazy na vlastní jména, kontextové překladově pro specifické významy heslového slova na základě kontextu konkrétního příkladu a variantní atribut pro glosy obsahující další pravopisné varianty heslového slova. Etymologické struktury byly rozlišeny na struktury kognátové pro jazyky současné vůči staré angličtině, reflexivní struktury pro formy slova v pozdějších etapách anglického jazyka a etymonové struktury v případě, že byl zmíněn zdrojový jazyk daného staroanglického slova.

Poslední kapitola sloužila jako seznam drobných změn pro stávající XML formát a dokument XSLT, které by zlepšily čitelnost webové aplikace a zároveň nebyly tak časově náročné na aplikaci jako značení složitých struktur popsané v předchozích kapitolách. Tyto změny spočívaly v odlišném grafickém odlišení různých elementů ve chvíli, kdy jsou vnořeny do jiného elementu, a v návrhu nového značení významové kategorie, které by umožňovalo odkazy jak na jednotlivé významy hesla, tak i odkazy na heslo jako takové, na rozdíl od současného značení, které umožňuje pouze odkazy na heslo.

## 12. Bibliography

- Atkins, S. B. T., & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.
- Baker, P. (2003). Toller at school: Joseph Bosworth, T. Northcote Toller and the progress of Old English lexicography in the nineteenth century. *Bulletin of the John Rylands Library*, 85(1), 95-114. <https://doi.org/10.7227/BJRL.85.1.6>
- Baker, P. S. (2012). *Introduction to Old English* (3rd ed.). Wiley-Blackwell.
- Bierbaumer, P., et al. (Eds.). (2007–2009). *Dictionary of Old English plant names*. <http://oldenglish-plantnames.org>
- Bosworth, J., & Toller, T. N. (1898). *An Anglo-Saxon dictionary*. Clarendon Press.
- Bosworth, J., & Toller, N. T. (1921). *An Anglo-Saxon dictionary supplement*. Clarendon Press.
- Cameron, A., et al. (Eds.). (1986–). *Dictionary of Old English*. University of Toronto. <https://doe.artsci.utoronto.ca/>
- Campbell, A. (1959). *Old English grammar*. Oxford University Press.
- Conway, P. (1989). Archival preservation: Definitions for improving education and training. *Restaurator: International Journal for the Preservation of Library and Archival Material*, 10(2), 47-60. <https://doi.org/10.1515/rest.1989.10.2.47>
- Conway, P. (2000). Overview: Rationale for digitization and preservation. In *Handbook for Digital Projects*. NEDCC.
- Crist, S. (2001). *Germanic Lexicon Project*. <http://www.germanic-lexicon-project.org/>
- Ellis, M. (1993). Old English lexicography and the problem of headword spelling. *ANQ: A Quarterly Journal of Short Articles, Notes and Reviews*, 6(1), 3-11.
- Garnett, J. M. (1898). [Review of *An Anglo-Saxon dictionary*, by T. N. Toller, J. R. Clark Hall, & H. Sweet]. *The American Journal of Philology*, 19(3), 323–328. <https://doi.org/10.2307/287977>
- Gradillas, M., & Llewellyn, D. W. T. (2023). Distinguishing digitization and digitalization: A systematic review and conceptual framework. *Journal of Product Innovation Management*, 1–32. <https://doi.org/10.1111/jpim.12690>
- Hall, J. R. C. (1894). *A concise Anglo-Saxon dictionary*. Swan Sonnenschein & Co.
- Jenkyns, J. (1991). The Toronto Dictionary of Old English resources: A user's view [Review of *Dictionary of Old English*]. *The Review of English Studies*, 42(167), 380–416. <http://www.jstor.org/stable/518350>
- Lee, K. H., Slattery, O., Lu, R., Tang, X., & McCrary, V. (2002). The state of the art and practice in digital preservation. *Journal of Research of the National Institute of Standards and Technology*, 107(1), 93-106. <https://doi.org/10.6028/jres.107.010>
- Lewis, R. E., et al. (Eds.). (2000–2018). *Middle English dictionary* (Online ed.). University of Michigan Press. <http://quod.lib.umich.edu/m/middle-english-dictionary/>
- Page, R. I. (1975). [Review of *An Anglo-Saxon dictionary based on the manuscript collections of Joseph Bosworth. Enlarged Addenda and Corrigenda*, by A. Campbell & T. N. Toller]. *Medium Ævum*, 44(1/2), 65–68. <https://doi.org/10.2307/43628077>
- Platt, J. (1884). The Bosworth-Toller Anglo-Saxon dictionary. *Transactions of the Philological Society*, 19(1), 237-246.
- Rothenberg, J. (2000). An experiment in using emulation to preserve digital publications.
- Schlutter, O. B. (1919). Some very pertinent remarks on Toller's supplement to Bosworth-Toller's *Anglo-Saxon dictionary*. *The Journal of English and Germanic Philology*, 18(1), 137–143. <http://www.jstor.org/stable/27700924>

- Swanepoel, P. (2003). Dictionary typologies: A pragmatic approach. In P. Van Sterkenburg (Ed.), *A practical guide to lexicography* (pp. [pages]). John Benjamins Publishing Company.
- Sweet, H. (1896). *The student's dictionary of Anglo-Saxon*. Oxford University Press.
- Tasovac, T. (2017). Toma Tasovac: TEI Lex - Toward a baseline encoding for legacy dictionaries [Video]. *YouTube*.  
<https://www.youtube.com/watch?v=JCdXSEzKwMo>
- Tasovac, T., & Romary, L., et al. (2018). TEI Lex-0: A baseline encoding for lexicographic data (Version 0.9.3). DARIAH Working Group on Lexical Resources. <https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html>
- TEI Consortium. *TEI P5: Guidelines for electronic text encoding and interchange*.  
<http://www.tei-c.org/Guidelines/P5/>
- Tichý, O. (2007). Digitization of old and middle English dictionaries. [Diploma thesis]. Faculty of Arts, Charles University.  
[https://dspace.cuni.cz/bitstream/handle/20.500.11956/9066/DPTX\\_2006\\_1\\_11210\\_ASZK10001\\_128223\\_0\\_27267.pdf](https://dspace.cuni.cz/bitstream/handle/20.500.11956/9066/DPTX_2006_1_11210_ASZK10001_128223_0_27267.pdf)
- Tichý, O., & Roček, M. (unpublished). Bosworth-Toller's *Anglo-Saxon dictionary online*.
- Tichý, O., Sean, C., & Toller, N. T. (2014–). *An Anglo-Saxon dictionary online*. Faculty of Arts, Charles University. <https://bosworthtoller.com/>
- Tichý, O., et al. (2021). Bosworth-Toller's *Anglo-Saxon dictionary online*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.  
<http://hdl.handle.net/11234/1-3532>
- Wright, J. (1914). *Old English grammar* (2nd ed.). Oxford University Press.
- Zgusta, L. (1971). *Manual of lexicography*. Academia.