ride. A review journal for digital editions and resources

published by the IDE

Review of “ShakespearePlaysPlus Text Corpus”

Shakespeare Corpus: ShakespearePlaysPlus, Mike Scott (ed.), 2006. http://lexically.net/wordsmith/support/shakespeare.html (Last Accessed: 03.07.2018). Reviewed by Katharina Mahler (CCeH, University of Cologne), english.textworks (at) gmail.com.

Abstract

ShakespearePlaysPlus is a freely available digital text corpus of William Shakespeare’s plays. The 37 plays were compiled from the Oxford University Press 1916 Edition of “The Complete Works of William Shakespeare” and annotated by Mike Scott for his own research in 2006. The plays are organized in three categories according to their type, i.e., comedies, historical plays and tragedies. The speeches of all characters have been extracted into separate text files. The text files are marked up in a pseudo-XML style and stored in Unicode. The corpus is downloadable as an extractable zip-file. This review presents a detailed look at the text corpus, its creation and composition. *ShakespearePlaysPlus* is compact and marked up with essential information, making it a durable, portable and easily reusable resource.

Introduction

1 *ShakespearePlaysPlus* is a freely available digital text corpus of William Shakespeare’s plays. The original source is the Oxford University Press 1916 Edition of “The Complete Works of William Shakespeare (The Oxford Shakespeare / OUP)”.[1] The corpus was compiled, processed and annotated by Dr. Mike Scott[2] for his own research

in 2006. The source texts were collected from the Online Library of Liberty (OLL) in digitized form, are in the public domain and “may be used freely for educational and academic purposes”. The corpus is available for download and re-use as a zip-file on the WordSmith Tools website[3] under “Extra downloads for WordSmith Tools”. Information concerning the origin of the text corpus, its design, contents and format is presented on the *ShakespearePlaysPlus* download page.

Aims, content and design of the corpus

2 This corpus provides a complete digital text collection of all of Shakespeare’s plays in standardized modern spelling, in a durable format that can easily be reused for further research. It can be used for analysis and research within the WordSmith environment or with any other lexical analysis tool. This primary source collection can be understood as a ‘general purpose corpus’, open to many possibilities for reuse and adaptation depending on the focus and interest of the user, ranging from literary to linguistic and quantitative as well as qualitative research.

3 The description of the content presented on the Wordsmith website is straightforward: “You get 37 plays, plus all the speeches of all the characters. That is, you get the whole play Hamlet, plus separately all the speeches of Prince Hamlet, all the speeches of Horatio, etc. There is also a list of the plays and their dates.” The overview of all the plays and years of publication is presented as a link to a new page.[4] The years of publication do not coincide with the dates presented on the OLL website, matching better with other Shakespeare sources.[5]

4 Basic relevant documentary information regarding format and design is presented in a pop-up link on the same page. The text files are stored in 16-bit Unicode format.[6] The zip-file is 5.4 MB large and unzips to 24.3 MB on the computer. It contains 41 file folders and 1387 files altogether.

5 The composition of this collection is guided by the principle of completeness, i.e., the complete plays of Shakespeare.[7] The 37 plays provided in the 1916 Oxford edition are organized according to three main categories, namely comedies, historical and tragedies. The plays and the respective “character speeches” sub-folders are available at the root level of each category folder.

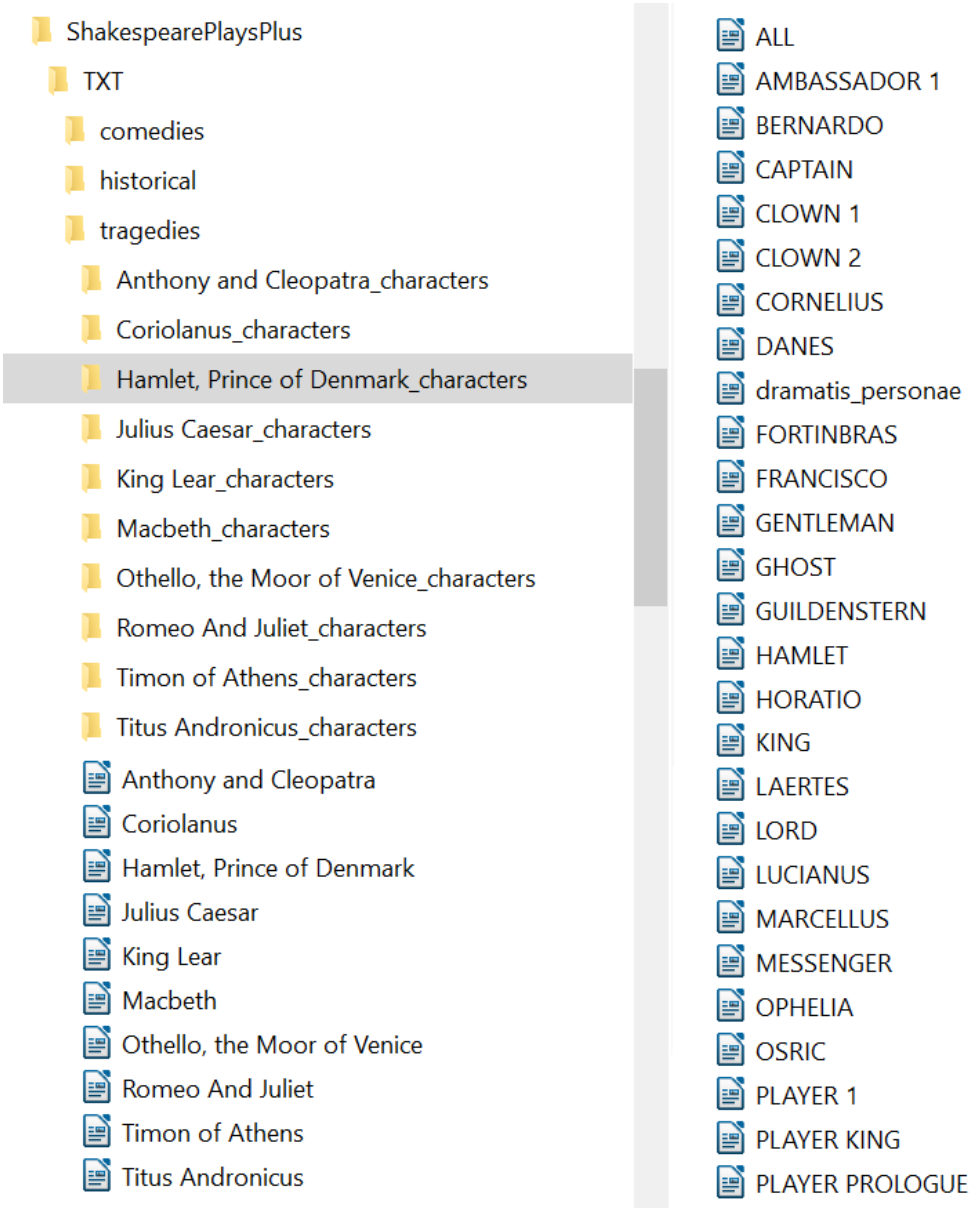


Fig. 1: Character speech folders, play files and *Hamlet*'s character speech files.

6 The 1313 extracted speech files are listed alphabetically in separate, accordingly named, sub-folders, e.g., "Hamlet, Prince of Denmark_characters" (see Fig. 1). A plain text alphabetical listing of each play's characters is included in the character speech folder, i.e., "dramatis_personae", providing an overview of the play's characters. These files reflect the *Dramatis Personae* listing at the beginning of each printed original play, however, they differ a bit from the source version in that they do not include further character descriptions (e.g., Friend to Hamlet, a Soldier, Officers, Courtiers, etc.) and some of the characters are listed more specifically, e.g., Clown 1 and Clown 2 instead of 'two clowns' as in the OUP.

```
ALL
AMBASSADOR 1
BERNARDO
CAPTAIN
CLOWN 1
CLOWN 2
CORNELIUS
DANES
FORTINBRAS
[…]
```

Code 1: Excerpt from *Hamlet*'s dramatis personae file list.

The dramatis personae file can be useful for specifying parameters in text analysis tools, e.g., the analysis of selected character speeches and their context, or creating concordances of a full play text without focus on the speakers.

7 The contents of the corpus provide what they promise, as described on the website. The structure of the data storage is clear and easy to understand. The simple structure makes it easy to integrate and reuse the text files in other systems.

Data modeling

8 The text files include embedded annotations in the form of "pseudo-XML angle-bracketed tags".[8] The annotations reflect the structure and contents of the play. Further analytic information not included in the HTML source texts in this form has been added, namely percentages indicating the relative position of the text segment[9] within the entirety of the play.[10] The percentages are based on the number of tokens altogether.

9 Each play consists of a single text file. Following the basic markup structure of HTML/XML,[11] each act, scene and character speech, as well as extra angle-bracketed stage directions, are surrounded by opening and closing angle-bracketed tags. Speeches are indented, making them easier to read while maintaining the visual form of standard HTML/XML. Each speech tag also contains the bracketed percentage information within the first line.

```
<ACT 1>
<SCENE 1>
```

```
<Elsinore. A Platform before the Castle.>
<STAGE DIR>
<Francisco at his post. Enter to him Bernardo.>
</STAGE DIR>
<BERNARDO> <0%>
Who's there?
</BERNARDO>
<FRANCISCO> <0%>
Nay, answer me; stand, and unfold yourself.
</FRANCISCO>
<BERNARDO> <0%>
Long live the king!
</BERNARDO>
[…]
</SCENE 1>
```

Code 2: Excerpt from the play file at the beginning of *Hamlet*.

10 In comparison, the character speech files are annotated without opening and closing brackets, deviating from the pseudo-XML style. The annotation at the beginning of each segment consists of the angle-bracketed number of the speech, act and scene, as well as the relative percentage information regarding the position within the play, as can be seen in the following:

```
<SPEECH 131><ACT 3><SCENE 1><42%>
To be, or not to be: that is the question:
Whether 'tis nobler in the mind to suffer
The slings and arrows of outrageous fortune,
Or to take arms against a sea of troubles,
And by opposing end them? To die: to sleep;
No more; and, by a sleep to say we end
The heart-ache and the thousand natural shocks
That flesh is heir to, 'tis a consummation
```

```
Devoutly to be wish'd. To die, to sleep;
To sleep: perchance to dream: ay, there's the rub;
```

Code 3: Excerpt from *Hamlet*'s speech file.

11 Overall, the data model is simple and easy to understand; however, the style of markup would make certain conversions necessary for reuse with XML-technologies. It is interesting that the creator specifically chose to model the texts in this manner, i.e., that the XML-version created during the transformation process was not used (see below). But considering that the corpus was originally created for personal research in WordSmith, by default ignoring angle bracket pairs and their content,[12] this markup style serves perfectly well. This markup style privileges speeches,[13] i.e., blending out all the lines starting with angle brackets leads to a text consisting only of the speeches.

Data acquisition and transformation

12 Besides essential information, i.e., the original book edition, the source of the digital texts and the final encoding, not much is told on the Wordsmith website regarding the creation and format of the original data and the transformation process behind this corpus. Upon my request, Mike Scott very kindly provided more detailed information, which is reused here with permission.

13 The 37 plays were downloaded individually in HTML, having "the merit of having detailed style information", as can be seen below:

```
<div class="sp"><span class="ital_speaker">Lys.</span>
<p style="margin-top: -0.5em;">I am, my lord, as well
deriv&#8217;d as he,</p>
<p class="p-no-indent1">As well possess&#8217;d; my love is
more than his;<span class="milestone_right"
title="Craig1916_line_100">100</span></p>
<p class="p-no-indent1">My fortunes every way as fairly
rank&#8217;d</p>
[…]
<div>
```

Code 4: Sample excerpt of source HTML from OLL, *A Midsummer Night's Dream*.

The original markup thus distinguishes the name of the character speaking from the lines spoken (`class="ital_speaker"`) and marks the line numbers in the OUP edition (`"milestone_right"`).

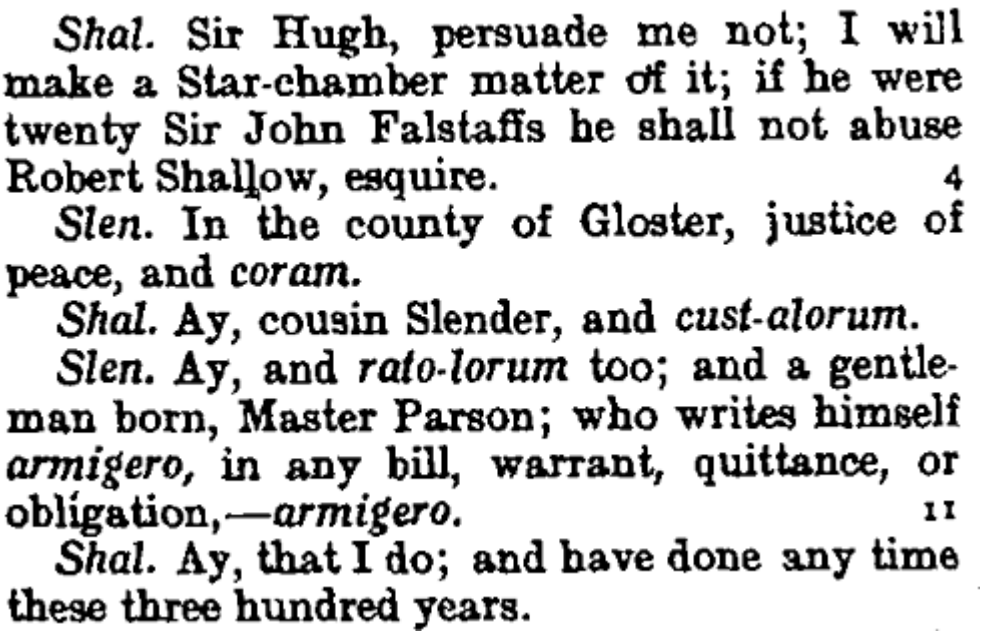
14 The HTML versions were converted to XML using Dreamweaver MX 2004. A program was written to convert the XML into plain text (TXT) with the desired markup. For this, it was necessary to

1. remove the standard headers and transform the text into Unicode with easily read punctuation instead of markup such as “`deriv&#8217;d`” (deriv’d).
2. find the XML markup for Dramatis Personae and build a standardised list of characters. This was problematic since in the text there are minor variations of spelling and it was not always easy to identify the character. In the example above, “Lys.” is easily identified as Lysander, but in some cases character names were simply absent from the Dramatis Personae, stage directions and the speech-wording.
3. identify Act and Scene numbers, mark up speech beginnings and endings so that these would be easy to locate, and to export all the speeches by all the characters in the plays to separate new files.

The final output is shown below:

```
<LYSANDER> <5%>
I am, my lord, as well deriv'd as he,
As well possess'd; my love is more than his;
My fortunes every way as fairly rank'd
[...]
</LYSANDER>
```

Code 5: TXT Extract of Fig. 5 from *A Midsummer Night’s dream*, play file.



Shal. Sir Hugh, persuade me not; I will make a Star-chamber matter of it; if he were twenty Sir John Falstaffs he shall not abuse Robert Shallow, esquire. 4

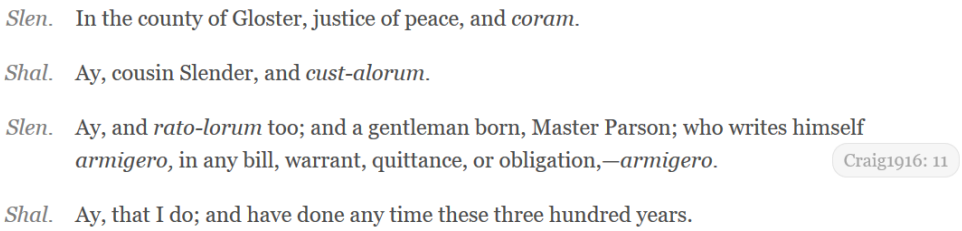
Slen. In the county of Gloster, justice of peace, and *coram.*

Shal. Ay, cousin Slender, and *cust-alorum.*

Slen. Ay, and *rato-lorum* too; and a gentleman born, Master Parson; who writes himself *armigero,* in any bill, warrant, quittance, or obligation,—*armigero.* 11

Shal. Ay, that I do; and have done any time these three hundred years.

Fig. 2: Lines 1 to 13 of *The Merry Wives of Windsor* from OUP, with line numbering and italics.



Slen. In the county of Gloster, justice of peace, and *coram.*

Shal. Ay, cousin Slender, and *cust-alorum.*

Slen. Ay, and *rato-lorum* too; and a gentleman born, Master Parson; who writes himself *armigero,* in any bill, warrant, quittance, or obligation,—*armigero.* Craig1916: 11

Shal. Ay, that I do; and have done any time these three hundred years.

Fig. 3: Lines 5 to 13 of OUP in HTML presentation, italics and line numbering as in OUP.

15 As can be seen above, some information was dropped from the OLL markup during the transformation into TXT format, i.e., the line numbering of the OUP and the markup reflecting the OUP pages (e.g., [52] for p. 52 of OUP). The cursive writing in the speeches is also not marked up in the TXT versions, as can be compared in the following examples (all from *The Merry Wives of Windsor*, see Fig. 2, 3, 4 and Code 6, 7).[14]

```
<SLENDER> <1%>
In the county of Gloster, justice of peace, and coram.
</SLENDER>
<SHALLOW> <1%>
Ay, cousin Slender, and cust-alorum.
</SHALLOW>
```

```
<SLENDER> <1%>
Ay, and rato-lorum too; and a gentleman born, Master
Parson; who writes himself armigero, in any bill,
warrant, quittance, or obligation,—armigero.
</SLENDER>
```

Code 6: Lines 5 to 13 of OUP in *ShakespearePlaysPlus*, without line numbering and italics.

Shal. Well, let us see honest Master Page. Is Falstaff there? 68

Eva. Shall I tell you a lie? I do despise a liar as I do despise one that is false; or as I despise one that is not true. The knight, Sir John, is there; and, I beseech you, be ruled by your well-willers. I will peat the door for Master Page. [*Knocks.*] What, hoa! Got pless your house here!

Page. [*Within*] Who's there? 76

Eva Here is Got's plessing, and your friend. and Justice Shallow; and here young Master Slender, that peradventures shall tell you another tale, if matters grow to your likings. 80

Fig. 4: Stage directions within speeches in square brackets, OUP.

16 In the OUP, stage directions within speeches are marked with square brackets and italics, as in Fig. 4. These are visually reproduced within the OLL and marked up in the HTML.

17 In the play files of *ShakespearePlaysPlus*, the target markup replaces the square brackets with angle brackets and surrounds these with <STAGE DIR> tags, as below.

```
<EVANS> <2%>
Shall I tell you a lie? I do despise a har as I do
despise one that is false; or as I despise one that is not
true. The knight, Sir John, is there; and, I beseech you, be
ruled by your well-willers. I will peat the door for Master
Page.
```

```
<STAGE DIR>
<Knocks.>
</STAGE DIR> What, hoa! Got pless your house here!
</EVANS>
```

Code 7: Stage directions within speeches in angle brackets in TXT file.

18 Some deviations – which have been corrected – occurred during the original transformation, however. The more relevant ones concerned 13 speech segments where in-speech stage directions resulted in lines of speech text being surrounded by angled brackets, enclosing 32 speech lines altogether, e.g.:

```
<BAPTISTA> <35%>
A mighty man of Pisa; by report
I know him well: you are very welcome, sir.
<STAGE DIR>
<To Hortensio.] Take you the lute, [To Lucentio.>
</STAGE DIR> and you the set of books;
You shall go see your pupils presently.
Holla, within!
```

Code 8: Speech mixed with stage directions in angle brackets in TXT file.[16]

19 In the single speech files, the 32 enclosed speech lines of the play files did not show up at all, thus deviating from the original OUP content and resulting in a loss of information.[17] Mike Scott set about correcting these deviations as soon as I informed him of my findings. The corrected version is now online for future downloads.[18]

Metadata

20 *ShakespearePlaysPlus* includes embedded metadata at the beginning of each play's text file. The metadata provide basic information about the original source of the texts as well as information about the collection at hand and the file format:

```
< Shakespeare - - HAMLET, PRINCE OF DENMARK >
< from Online Library of Liberty (http://oll.libertyfund.org) >
< Unicode .txt version by Mike Scott (http://www.lexically.net) >
< from "The Complete Works of William Shakespeare" >
```

```
< ed. with a glossary by W.J. Craig M.A. >
< (London: Oxford University Press, 1916) >
```

Code 9: Metadata from the *Hamlet* play file.

The metadata are not repeated in the single characters' speech files. A date of creation is not provided, but the entire corpus was created in 2006.[19]

Further aspects: access, preservation, reuse

21 As the basic format of the text collection is 16-bit Unicode,[20] it is a durable corpus[21] that is readable, processible and accessible for the long term. No specific user interface is provided or necessary. This corpus consists of plain text files, so a basic editor or text program can in principle suffice for human reading. For specific research questions, WordSmith or any other tool that works with annotated text files can be used as an interface. Whether the format or markup have to be adapted depends on the environment, e.g., for use with XML-technologies, the markup would have to be adapted to be well-formed (see footnote 11).[22]

22 The text collection has brief documentation regarding source, format, annotation and design on the website. The original source of the collection, as well as creator, contact and download website address, are provided in the metadata of each of the plays. There are no persistent identifiers. The original source texts are in the public domain and no specific license is in use.

23 Due to its simple and durable design, this text corpus has good prospects of being available for research and reuse for a long time. Continuous access is provided by the WordSmith Tools Website. The website entered its 22th year of existence,[23] so one may be optimistic that it will remain online well into the future. The corpus is not officially mirrored or archived in an academic repository, but it is archived at the WayBackMachine Internet Repository.

24 Regarding reuse, at least one author has used this corpus as a basis for her MA thesis investigating key word clusters in the dialogue of Shakespeare's female characters.[24] Scott does not have additional records of people or projects using this corpus, but expects that "quite a lot of students will have downloaded it for their term papers".[25]

Conclusion

25 *ShakespearePlaysPlus* is a practical text collection that is open for many potential reuses. Important aspects behind the creation of the corpus were completeness, durability, reusability and standardized modern English spelling. The markup reflects the structure of the play (act, scene, stage direction, speeches of characters), and the relative position of the speech tags. Considering more complex data models and current standards of research data management, some modifications could possibly be undertaken, e.g., adapting the metadata (adding distinct IDs/DOIs and creation dates, including the metadata in the speech files) and the markup to be well-formed XML.

26 There is no particular theoretical stance behind the text corpus, at least not explicitly, but a special interest in the analysis of speeches may be presumed. As Mike Scott writes on the website, “I made this version for my own research. If you use it please let me know!” And further in personal correspondence, “I thought others might find it useful, no other great ambition!”.

27 *ShakespearePlaysPlus* is a durable and portable text corpus marked up with relevant information. Making this practical corpus created for personal use available to the public for further reuse and research was a collegial and friendly deed in the spirit of open access, adding to the pool of digital resources available for linguistic and literary Shakespeare research.

Notes

1. Original source of the online version: William Shakespeare, *The Complete Works of William Shakespeare (The Oxford Shakespeare)*, ed. with a glossary by W. J. Craig M. A. (Oxford University Press, 1916). See https://web.archive.org/web/20180426191653/http://oll.libertyfund.org/titles/shakespeare-the-complete-works-of-william-shakespeare-the-oxford-shakespeare for the online OLL version. Several formats are available for download, including HTML (“Every effort has been taken to translate the unique features of the printed book into the HTML medium.“), Ebook and ePub, as well as a facsimile of the original book as an image-based PDF.

2. Dr. Mike Scott, Liverpool University (1990-2009), Aston University (2010-), see https://web.archive.org/web/20180126155259/http://www.aston.ac.uk/lss/staff-directory/

scottmdr/. Developer of WordSmith Tools (1996-2018). For further publications, see https://web.archive.org/web/20180426191409/http://lexically.net/publications/publications.htm.

3. Available on the Wordsmith Homepage https://web.archive.org/web/20180703103234/http://www.lexically.net/wordsmith/index.html under “Downloads”, then “Extras”, see https://web.archive.org/web/20180118112105/http://lexically.net/wordsmith/support/shakespeare.html.

4. See https://web.archive.org/web/20180703103115/http://lexically.net/downloads/corpus_linguistics/shakespeare_plays_dated_plain.txt.

5. For simple accessibility, only freely available online sources are cited in this review. Please refer to the canon of Shakespeare research for in-depth discussion. The publication dates of Shakespeare’s plays are not precisely historically documented, a topic which will not be discussed here in length. See, e.g., the timelines presented online by the Royal Shakespeare Company https://web.archive.org/web/20180118115432/https://www.rsc.org.uk/shakespeares-plays/timeline, on Open Source Shakespeare https://web.archive.org/web/20180118115735/https://www.opensourceshakespeare.org/views/plays/plays_date.php, or Shakespeare Online https://web.archive.org/web/20180118120823/http://shakespeare-online.com/keydates/playchron.html, which also lists the presumable years of first performance vs. publication.

6. There are no special characters that make 16-bit Unicode necessary, but, as Scott notes, it is a generally useful encoding.

7. The number of plays written by Shakespeare is a controversial topic. The Royal Shakespeare Company (RSC), e.g., writes “[i]t is believed that he wrote around 38 plays, including collaborations with other writers”, see https://web.archive.org/web/20180118115432/https://www.rsc.org.uk/shakespeares-plays/timeline. The RSC include the play “Two noble Kinsmen” (1613-1614) in their listing, which is not included in the Oxford 1916 edition. This play is also listed on Shakespeare Online https://web.archive.org/web/20180124122014/http://www.shakespeare-online.com/keydates/playchron.html, where it is noted that “all but a few scholars believe it not to be an original work by Shakespeare. The majority of the play was probably written by John Fletcher, who was a prominent actor and Shakespeare's close friend” (Mabillard 2000).

8. See link ‘Format information’ on the download page, https://web.archive.org/web/20180118112105/http://lexically.net/wordsmith/support/shakespeare.html.

9. The OLL HTML does include information about the numbering of the text lines, but not in percentage form. For example, one sees Craig1916: 4 in the front-end browser version, standing for line 4, reflecting the numbering as printed in the original book, see, e.g., https://web.archive.org/web/20180124112122/http://oll.libertyfund.org/titles/shakespeare-hamlet-prince-of-denmark--5. In the HTML code, each text line has a specific embedded numbering (as viewed in December 2017).

10. At the very beginning of a play it is <0%> and at the end of the play <100%> (or <99%>, if it ends with a single long speech, e.g., in *The Life of King Henry V*). Depending on the length of the speeches, a single percent may be repeated within the markup of several following character speeches (see Fig. 3), or one long speech can cover more than a single percent, e.g., in *The Life of King Henry V*, a King Henry speech is marked with <5%>, and the following speech in marked with <7%> (Act 1, Scene 2).

11. The format is, however, not well-formed XML because (1) the tag names contain whitespace, (2) angle brackets are used to surround the stage directions, percentages, and metadata, (3) metadata, percentages, and markup in character speech files are not surrounded by opening and closing tags, i.e., not properly nested and (4) there is no root element.

12. See http://www.lexically.net/wordsmith/Handling%20BNC/tag_file.htm, last accessed May 14, 2018. Tags can be retained for the analysis with WordSmith when a special tag file is created.

13. Compared to, e.g., printed text from the source edition.

14. Screenshots from the fascsimile of the OUP (p. 51 of OUP, p. 62 of 1390 in PDF), see https://web.archive.org/web/20180428131356/http://lf-oll.s3.amazonaws.com/titles/1608/0612_Bk_Sm.pdf.

16. Excerpt from *The Taming of the Shrew*, Baptista at 35%. Depending on the setting of the research environment, angle-bracketed content may be ignored (as, e.g., by default in WordSmith), thus excluding these 32 lines from analysis.

17. Compared to the number of lines in Shakespeare's plays altogether, the impact of the 32 enclosed speech lines in the play files and the 32 corresponding missing lines in the

speech files may be deemed quite minimal — yet, depending on the research focus, they could have distorted results to a certain degree. For example, in Rosalind's <speech 192> at 89 % (from *As You Like It*), entire nine lines of speech were missing.

18. As of the 14th of June, 2018.

19. Personal correspondence.

20. For the Unicode Character Encoding Stability Policies, see https://web.archive.org/web/20180118134036/http://www.unicode.org/policies/stability_policy.html.

21. Regarding encoding choices, Folger Digital, e.g., notes that they offer TXT versions in ASCII-7, a subset of Unicode, and that these "files are the most likely to render properly in the widest number of applications and the least likely to present conversion errors when being incorporated into text analysis tools." Their unadorned TXT files are "for projects and applications where simplicity and/or stability is the highest priority", see https://web.archive.org/web/20180118123036/http://www.folgerdigitaltexts.org/download/txt.html.

22. It may be noted that other online Shakespeare sources, e.g., Folger Digital Texts, provide XML and TEI markup for download. Folger Digital provides complex XML markup, including markup for line breaks, individual words and punctuation marks, as well as IDs for each tag. The TEISimple version is also quite complex, including linguistic annotation and single line IDs, see https://web.archive.org/web/20180124140232/http://www.folgerdigitaltexts.org/download/. Christof Schöch (2014) notes in his review of Folger Digital that the texts were marked up in an exemplary manner for maximal interoperability with other resources.

23. See https://web.archive.org/web/20180703103234/http://www.lexically.net/wordsmith/index.html.

24. Demmen, Jane (2009). *Charmed and chattering tongues: Investigating the functions and effects of key word clusters in the dialogue of Shakespeare's female characters*. Lancaster University, MA thesis (unpublished). See https://web.archive.org/web/20180118142818/http://lexically.net/wordsmith/corpus_linguistics_links/Jane%20Demmen_MA_word%20clusters_in_Shakespeare%27s%20plays.pdf.

25. Personal correspondence.

References

Folger Digital Texts. “Download.” https://web.archive.org/web/20180124140232/http://www.folgerdigitaltexts.org/download/.

Mabillard, Amanda. 2000. “The Chronology of Shakespeare's Plays.” *Shakespeare Online*. https://web.archive.org/web/20180124122014/http://www.shakespeare-online.com/keydates/playchron.html.

Online Library of Liberty (OLL), https://web.archive.org/web/20180118112729/http://oll.libertyfund.org.

Open Source Shakespeare. 2003-2018. “Shakespeare's plays, listed by presumed date of composition.” https://web.archive.org/web/20180118115735/https://www.opensourceshakespeare.org/views/plays/plays_date.php.

Royal Shakespeare Company. “Timeline of Shakespeare's plays. A chronological list of Shakespeare's plays by decade.” https://web.archive.org/web/20180118115432/https://www.rsc.org.uk/shakespeares-plays/timeline.

Schöch, Christof. 2014. “Michael Poston and Rebecca Niles, eds.: Folger Digital Texts. Washington: The Folger Shakespeare Library, 2012-2013.” *Variants - The Journal of the European Society for Textual Scholarship*, pp. 16-20. https://zenodo.org/record/13745.

Scott, Mike. 2016. “Extra Downloads for WordSmith Tools: Shakespeare Corpus.” *WordSmith Tools.* https://web.archive.org/web/20180118112105/http://lexically.net/wordsmith/support/shakespeare.html.

Shakespeare Online. 1999-2012. “The Chronology of Shakespeare's Plays.” https://web.archive.org/web/20180118120823/http://shakespeare-online.com/keydates/playchron.html.

Shakespeare, William. 1916. *The Complete Works of William Shakespeare (The Oxford Shakespeare)*. Edited with a glossary by W. J. Craig M. A. Oxford: Oxford University Press. HTML: https://web.archive.org/web/20180126162402/http://oll.libertyfund.org/titles/shakespeare-the-complete-works-of-william-shakespeare-the-oxford-shakespeare. Facsimile: https://web.archive.org/web/20180428131356/http://lf-oll.s3.amazonaws.com/titles/1608/0612_Bk_Sm.pdf.

WordSmith Tools. Lexical Analysis Software and Oxford University Press. https://web.archive.org/web/20180703103234/http://www.lexically.net/wordsmith/index.html.

Factsheet

Resource reviewed	
Title	Shakespeare Corpus: ShakespearePlaysPlus
Editors	Mike Scott
URI	http://lexically.net/wordsmith/support/shakespeare.html
Publication Date	2006
Date of last access	03.07.2018

Reviewer	
Surname	Mahler
First Name	Katharina
Organization	CCeH, University of Cologne
Place	Cologne
Email	english.textworks (at) gmail.com

General Information		
Bibliographic description	Can the text collection be identified in terms similar to traditional bibliographic descriptions (title, responsible editors, institution, date(s) of publication, identifier/address)? (cf. Catalogue 1.1)	yes
Contributors	Are the contributors (editors, institutions, associates) of the project documented? (cf. Catalogue 1.3)	yes
Contacts	Is contact information given? (cf. Catalogue 1.4)	yes
Aims		
Documentation	Is there a description of the aims and contents of the text collection? (cf. Catalogue 2.1)	yes
Purpose	What is the purpose of the text collection? (cf. Catalogue 2.2)	Research, General purpose
Kind of research	What kind of research does the collection allow to conduct primarily? (cf. Catalogue 3.1.8)	Quantitative research

Self-classification	How does the text collection classify itself (e.g. in its title or documentation)? (cf. Catalogue 2.3)	Corpus
Field of research	To which field(s) of research does the text collection contribute? (cf. Catalogue 2.2)	Literary studies, Linguistics
Content		
Era	What era(s) do the texts belong to? (cf. Catalogue 2.5)	Classics, Early Modern
Language	What languages are the texts in? (cf. Catalogue 2.5)	English
Types of text	What kind of texts are in the collection? (cf. Catalogue 2.5)	Literary works
Additional information	What kind of information is published in addition to the texts? (cf. Catalogue 2.5)	none
Composition		
Documentation	Are the principles and decisions regarding the design of the text collection, its composition and the selection of texts documented? (cf. Catalogue 3.1.1-3.1.3)	yes
Selection	What selection criteria have been chosen for the text collection? (cf. Catalogue 3.1)	Author, Genre
Size		
Texts/records	How large is the text collection in number of texts/ records? (cf. Catalogue 3.1.4)	> 1000
Tokens	How large is the text collection in number of tokens? (cf. Catalogue 3.1.4)	100,000- 1 Mio.
Structure	Does the text collection have identifiable sub-collections or components? (cf. Catalogue 3.1.5)	yes
Data acquisition and integration		
Text recording	Does the text collection record or transcribe the textual data for the first time? (cf. Catalogue 3.1.6)	no
Text integration	What kind of material has been taken over from other sources? (cf. Catalogue 3.1.6)	Full texts, Annotations

Quality assurance	Has the quality of the data (transcriptions, metadata, annotations, etc.) been checked? (cf. Catalogue 3.1.7)	yes
Typology	Considering aims and methods of the text collection, how would you classify it further? For definitions please consider the help-texts. (cf. Catalogue 3.1.8)	Corpus, Canon, Complete works/ œuvre
Data Modelling		
Text treatment	How are the textual sources represented in the digital collection? (cf. Catalogue 3.2.1)	Normalized transcription
Basic format	In which basic format are the texts encoded? (cf. Catalogue 3.2.4)	Plain text
Annotations		
Annotation type	With what information are the texts further enriched? (cf. Catalogue 3.2.2)	Structural information
Annotation integration	How are the annotations linked to the texts themselves? (cf. Catalogue 3.2.2)	Embedded
Metadata		
Metadata type	What kind of metadata are included in the text collection? (cf. Catalogue 3.2.3)	Descriptive
Metadata level	On which level are the metadata included? (cf. Catalogue 3.2.2)	Individual texts
Data schemas and standards		
Schemas	What kind of data/metadata/annotation schemas are used for the text collection? (cf. Catalogue 3.2.4)	Project specific schema
Standards	Which standards for text encoding, metadata and annotation are used in the text collection? (cf. Catalogue 3.2.4)	other:
Provision		
Accessability of the basic data	Is the textual data accessible in a source format (e.g. XML, TXT)? (cf. Catalogue 4.1)	yes
Download	Can the entire raw data of the project be downloaded (as a whole)? (cf. Catalogue 4.2)	yes

Technical interfaces	Are there technical interfaces which allow the reuse of the data of the text collection in other contexts? (cf. Catalogue 4.2)	none
Analytical data	Besides the textual data, does the project provide analytical data (e.g. statistics) to download or harvest? (cf. Catalogue 4.3)	no
Reuse	Can you use the data with other tools useful for this kind of content? (cf. Catalogue 4.4)	yes
User Interface		
Interface provision	Does the text collection have a dedicated user interface designed for the collection at hand in which the texts of the collection are represented and/or in which the data is analyzable? (cf. Catalogue 5.1)	no
Preservation		
Documentation	Does the text collection provide sufficient documentation about the project in general as well as about the aims, contents and methods of the text collection? (cf. Catalogue 6.1)	yes
Open Access	Is the text collection Open Access? (cf. Catalogue 6.2)	yes
Rights		
Declared	Are the rights to (re)use the content declared? (cf. Catalogue 6.2)	yes
License	Under what license are the contents released? (cf. Catalogue 6.2)	No explicit license / all rights reserved
Persistent identification and addressing	Are there persistent identifiers and an addressing system for the text collection and/or parts/objects of it and which mechanism is used to that end? (cf. Catalogue 6.3)	none
Citation	Does the text collection supply citation guidelines? (cf. Catalogue 6.3)	no
Archiving of the data	Does the documentation include information about the long term sustainability of the basic data (archiving of the data)? (cf. Catalogue 6.4)	no

Institutional curation	Does the project provide information about institutional support for the curation and sustainability of the project? (cf. Catalogue 6.4)	no
Completion	Is the text collection completed? (cf. Catalogue 6.4)	yes
Personnel		
Editors	Mike Scott	
Encoders	Mike Scott	
Administrators	Mike Scott	