Computer-Aided Linguistic Analysis for a Single Manuscript Witness: Preparing to Map the Opentext.org annotation

Smith, Catherine; O'Donnell, Matthew Brook

DOI:
10.1163/9789004335936_005

License:
Other (please provide link to licence statement

Document Version
Early version, also known as pre-print

Citation for published version (Harvard):
Smith, C & O'Donnell, MB 2016, Computer-Aided Linguistic Analysis for a Single Manuscript Witness: Preparing to Map the Opentext.org annotation. in L Dow, C Evans & A Pitts (eds), *The Language and Literature of the New Testament: Essays in Honour of Stanley E. Porter's 60th Birthday.* Biblical Interpretation Series, vol. 150, Brill, pp. 106-137. https://doi.org/10.1163/9789004335936_005

COMPUTER-AIDED LINGUISTIC ANALYSIS FOR A SINGLE MANUSCRIPT WITNESS: PREPARING TO MAP THE OPENTEXT.ORG ANNOTATION

Catherine Smith[1] and Matthew Brook O'Donnell

Abstract

Porter has been a strong advocate of adopting a single manuscript approach over the traditional eclectic text approach as the basis for analysing the New Testament. This has implications for exegetical work including the use of linguistic models, another dominant strand in his work. Existing electronic texts of the New Testament that include various levels and types of annotation are based on eclectic texts, most often one of the editions of the Nestle-Aland or United Bible Societies. This is true for the OpenText.org annotation that implements a multilevel word, phrase and clausal analysis on the basis of NA^{27}. In this paper we examine the issues of a text critical, methodological and technical nature that must be addressed if the OpenText.org syntactic annotation is to be successfully moved from Nestle-Aland to a single manuscript witness. Beginning with Codex Sinaiticus we use the Pauline corpus as a test case to illustrate the challenges and implications of this mapping and make some proposals for future annotation practice.

Introduction

> I would suggest that we recognize what tacitly is the case and move away from an idealized eclectic text that never existed in any Christian community back to the codices that still form the basis of our modern textual tradition.[2]

Porter is a strong advocate of adopting the text of a single manuscript as the basis for studying the New Testament rather than the more widely used critical texts. The core of Porter's argument is that our modern eclectic texts are heavily reliant on a few of the major codices, namely Codex Vaticanus and Codex Sinaiticus that formed the 'Neutral' text of Westcott and Hort in their edition of 1881 and were considered by the editors to be the closest to the original. This edition along with the 8th edition of Tischendorf

[1] Catherine Smith's research in the present paper has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 283302 (COMPAUL).

[2] Porter, "Papyrological Evidence," 176-8.

(1868-72), itself heavily influenced by his own discovery of Codex Sinaiticus in the years since the publication of his 7th edition, formed the basis of Nestle's first edition of 1898.[3] Porter notes that when the first edition of Westcott and Hort and the 8th edition of Tischendorf were published only one papyrus was known and only parts of that were published.[4] From the third edition onwards Nestle also used Weiss's edition but only seven papyri were known when that was published in 1900;[5] the running text of NA^{25} was estimated by the editors themselves to differ from the Westcott and Hort text in only 558 places and the first Nestle text in only 700.[6] Porter also quotes Robinson's calculation that the text of NA^{26} is 99.5 percent the same as the Westcott and Hort edition showing the limited impact of the discovery of the papyri and the continued reliance on a few major codices.[7] He notes that Petersen finds that there is nowhere in the critical apparatus of NA^{27} where a reading is supported on the basis of papyrus evidence or papyrus and patristic evidence alone[8] concluding that the papyri 'do not so much represent a text as support readings and push back in time readings found in the major codexes'.[9] Elsewhere Porter also argues that it is perhaps wise not to give too much weight to the papyri given how little we know of their original purpose and

[3] Porter, *Text, Transmission, Translation*, 74. For a more details discussion of the development of printed editions see Porter *Text, Transmission, Translation*, 36-51.

[4] Porter, *Text, Transmission, Translation*, 73.

[5] Porter, *Text, Transmission, Translation*, 73.

[6] Porter, *Text, Transmission, Translation*, 72.

[7] Robinson, "Appendix," 551 in Porter, *Text, Transmission, Translation*, 72.

[8] Petersen, "What Text," 138 in Porter, *Text, Transmission, Translation*, 73. In his response to Petersen at the SBL annual meeting in 1998 Parker cites a number of examples where the papyri may not be the only evidence cited but do seem to have influenced the reading present in the editorial text of NA^{27} these are Matt 20:30; Luke 11:11; Heb 11:37 and 1Pet 2:5. In each of these cases (excepting von Soden for 1 Peter) NA^{27} is the first of the critical editions to include these readings in the critical text: all with support from at least one papyrus. We are grateful to Professor Parker for providing a copy of his response.

[9] Porter, *Text, Transmission, Translation*, 73, after Petersen, "What Text," 138-9.

compositional character.[10] An additional reason to use a single manuscript as the basis of analysis is that our modern editions are eclectic and do not conform to any single manuscript. According to Porter this means that 'the critical text of the New Testament today is only as old as nineteenth-century scholarship'[11] and instead he suggests that 'those seeking the original text of the New Testament consider seeking it through individual manuscripts'.[12] Porter summarises his conclusions as follows:

> If Westcott and Hort's edition is clearly based on the two major codexes, and the current text is 99.5 percent the same—that is, with all of the other evidence that has been bought to bear, including papyri and all else, only 0.5 percent different—it seems as if we are already in essence using the text of the two major codexes. If our goal is to seek the earliest text that we legitimately can find, without abandoning the claim to be seeking the original even if we know that we can only get back so far, then it makes sense to use the earliest actual texts we can find… These actual texts were written and used in the early church, and in reality they get closer to the original autographs in terms of quantifiable evidence than a text edited in the nineteenth, twentieth, and now twenty-first centuries.[13]

Two other major strands of Porter's research are discourse analysis and sociolinguistics, fields in which the choice of text is of particular relevance. Two studies in this area, both focusing on the book of Acts, have addressed this particular question and support Porter's call to select a single manuscript witness. Read-Heimerdinger's work published 2002 looks at the contribution of discourse analysis to textual criticism, while Snyder's 2014 work takes a sociolinguistic approach to studying the construction of Christian identity in the texts with a particular focus on the social context. Along with the book of Acts Snyder's study also includes two apocryphal texts, the Acts of John and the Acts of Philip. Snyder reports that it was not until she began to study the two apocryphal texts, which survive in far fewer and far more divergent witnesses than the

[10] Porter, "Papyrological Evidence," 176.

[11] Porter, *Text, Transmission, Translation*, 74.

[12] Porter, *Text, Transmission, Translation*, 74.

[13] Porter, *Text, Transmission, Translation*, 75.

book of Acts, that she began to question her initial approach of using the Nestle-Aland text as the basis for her analysis of the New Testament text.[14] In light of being forced by the nature of the witnesses to select a single manuscript approach to the apocryphal text Snyder then followed the same path for the book of Acts and selected Codex Vaticanus as the basis for her analysis.[15] While Snyder freely admits that her choice of text and her selection of linguistic variables for analysis produced results that are no different that they would have been had she used the Nestle-Aland text[16] the theoretical point is still an important one. The crucial observation, which agrees with Porter's line of argument, is this: "one must generally choose a particular version to analyze, or have it chosen on one's behalf, because reconstructed texts of ActsAp [the book of Acts] such as the Nestle-Aland already privilege certain versions while disfavoring others".[17]

In the other of the two studies Read-Heimerdinger selects Codex Bezae as her text for analysis, one of the witnesses less favoured in critical texts, although like Snyder she admits that many linguistic studies of the book of Acts based on the Nestle-Aland text are not particularly problematic due to the strong influence of Alexandrian manuscripts in the editorial text.[18] With regards to the text of Codex Bezae in the book of Acts, Read-Heimerdinger argues that the lack of a consensus is due to it being studied in its context as an example of the disparate group of manuscripts classified under the Western text-type rather than as a text in its own right reinforcing the need to use single manuscripts as the basis for analysis.[19] Her work deals specifically with the application of techniques from discourse analysis to questions of textual criticism. This throws even

[14] Snyder, *Language and Identity*, 19.
[15] Snyder, *Language and Identity*, 36.
[16] Snyder, *Language and Identity*, 38.
[17] Snyder, *Language and Identity*, 20.
[18] Read-Heimerdinger, *Bezan Text*, 52.
[19] Read-Heimerdinger, *Bezan Text*, 5.

more focus on the choice of text. Her selection of linguistic features also deliberately targets areas that can contribute to the discussion of textual variation[20] and, she claims, often relate to areas which are traditionally seen as insignificant by textual critics and which frequently go unrecorded in modern eclectic texts.[21] Read-Heimerdinger concludes that "[i]n examining a text from the perspective of discourse analysis, It makes much more sense to work from the text of actual MSS, for the interest of a discourse analyst is precisely in elucidating rules from what has been said or written".[22] She also goes on to echo the quote from Porter at the start of this article saying that her analysis is "based on a comparison of the texts transmitted in early MSS that can be assumed to have been used by actual communities"[23] reflecting the concern of discourse analysis with the possible audiences of a text. The importance of Porter's suggestion then is certainly felt among those researchers interested in questions of sociolinguistics and discourse analysis. Snyder sums up the situation for sociolinguistic analysis as follows but it is just as applicable to discourse analysis and similar studies:

> If one uses an eclectic text to address sociolinguistic questions, one is essentially comparing words drawn from manuscripts of different dates and provenances, a strategy that only makes sense if one assumes that every individual whose words are reflected in the various manuscripts used language in the exact same way. Surely this is a dangerous assumption to make given how language changes over time and how the cultural background, geographical origin, and life experiences of individuals affect the way they speak. A sociolinguistic study of a modern reconstructed text could yield interesting insights into the linguistic sensibilities of modern scholars, but it would not necessarily elucidate the nuances of how ancient writers used words.[24]

When the decision to use a single manuscript witness as the text for analysis is

[20] Read-Heimerdinger, *Bezan Text*, 44-9.
[21] Read-Heimerdinger, *Bezan Text*, 42, 52.
[22] Read-Heimerdinger, *Bezan Text*, 52.
[23] Read-Heimerdinger, *Bezan Text*, 52.
[24] Snyder, *Language and Identity*, 20.

made several further questions arise, the most obvious being which manuscript to select. Codex Sinaiticus is the oldest surviving witness to the entire Greek New Testament and Porter suggests we use this as a replacement for modern critical editions.[25] He also acknowledges that for individual books one could use Codex Vaticanus, like Snyder, or a particular papyrus, such as P46 for the Pauline Epistles.[26] Since the OpenText.org annotation is available for the entire New Testament, this paper will follow Porter's suggestion and focus primarily on Codex Sinaiticus. Another question that arises is how the exact text of the selected manuscript should be determined. This is an issue that has not been addressed in detail by any advocates of this model. Snyder, Read-Heimerdinger and Porter all make the point that a single manuscript is still an eclectic text but an ancient eclectic text rather than a modern one.[27] Porter and Pitts go further in their introduction to New Testament textual criticism stating that this approach 'leaves the text-critical responsibilities with the ancients'[28] but in reality if a scholar choses to use a single manuscript as the basis of their analysis they should be prepared to take on the text-critical responsibilities themselves as described below. Ignoring the history of the transmission of the text represented in a manuscript can be as problematic as adopting a modern critical edition without any regard to the way that text has been created. You must become the editor of your own text.

Porter and Pitts do draw attention to some editorial decisions that must be made when they say '[w]hat it means to use a singular manuscript like Codex Sinaiticus as the

[25] Porter, *Text, Transmission, Translation*, 75.

[26] Porter, *Text, Transmission, Translation*, 75.

[27] Snyder, *Language and Identity*, 39; Read-Heimerdinger, *Bezan Text*, 5; Porter and Pitts, *Introduction*, 95.

[28] Porter and Pitts, *Introduction*, 95.

basis of a modern edition might vary in some people's minds (e.g., whether spelling was regularized or clear errors in the text were corrected)'.[29] Elsewhere Porter also states 'the various correcting hands of Sinaiticus attest to its use, but also raise a number of important critical issues'[30] without going any further into how these critical issues could be addressed. A certain amount of text-critical knowledge and maybe even paleographical skills would be needed to determine which spelling conventions should be regularised and to identify what are clear errors and decide how to correct them. The issue of the different hands present in a manuscript is a particularly important one in Codex Sinaiticus, one of the most heavily corrected biblical manuscripts we have.[31] Several layers of correction can be determined in Codex Sinaiticus and many hands can be identified consistently. The corrections range from those contemporary with the first hand to far later corrections stretching into the twelfth century.[32] Milne and Skeat in particular note that corrections made in the scriptorium are easily distinguishable from the later correctors.[33] In order to use the text of Codex Sinaiticus an editorial decision must be made as to which of these strands of correction is to be treated as the text of the manuscript. Since so much is known about the correction layers in Codex Sinaiticus for the purposes of this particular study we will take the text of Codex Sinaiticus to be the manuscript text as it looked when it left the scriptorium as far as that can be reconstructed. That is corrections by the first hand and the corrector known as S1 will replace the initial first hand text where they differ and any later corrections will be

[29] Porter and Pitts, *Introduction*, 95.
[30] Porter, *Codex Sinaiticus*.
[31] Wachtel, "The Corrected New Testament," 97.
[32] Wachtel, "The Corrected New Testament," 98.
[33] Milne and Skeat, *Scribes and Correctors*, 40.

ignored.[34] For other manuscripts about which little or nothing of their production process is known the decision of which text to select may be simpler but perhaps less accurate. The key thing when selecting both a text to use and, if a single manuscript text is selected the choice of text to extract, is that is should be a conscious choice and whatever is selected, the text, its creation, and perhaps transcription, are fully understood and handled appropriately.

The OpenText.org project[35] arose out of O'Donnell's doctoral research on the application of corpus linguistic methodologies to the study of the Greek New Testament[36] with Porter as a primary partner. The project aims to create a linguistically annotated corpus designed for the study of New Testament Greek with a particular focus on discourse analysis.[37] The annotation is based on a systemic-functional model and includes analysis at word, word group, clause and discourse levels. For ease of annotation each phase of the analysis was initially completed separately and stored in separate XML files. When the project started investigating search options these different layers were pulled into a single XML file for ease of processing. This results in a complex XML structure: one of the questions addressed in this study is whether or not this resulting structure is suitable for mapping to other base. Open-source and open data philosophies were central to the development of OpenText.org[38] so basing the analysis

[34] In the rest of this paper this particular textual strand will be referred to as the text of Codex Sinaiticus although it should be understood as the text we selected from the multiple texts attested in Codex Sinaiticus.
[35] http://www.opentext.org
[36] O'Donnell, *Corpus Linguistics*.
[37] For a discussion of the design of the corpus see O'Donnell, *Corpus Linguistics*, 102-37 and Porter and O'Donnell, 'Theoretical Issues'; for an introduction to the annotation model which is discussed further below see O'Donnell, *Corpus Linguistics*, 168-201 and O'Donnell, *Introducing*.
[38] Porter, *Linguistic Analysis*, 39-46, in particular 45-6.

on texts that can be further distributed along with the annotation is an important factor when considering a change in text.

The original plan for OpenText.org was to use the text of Codex Sinaiticus as the text for annotation but the text of Codex Sinaiticus was not available in electronic form and several modern editions were.[39] Indeed one of the main reasons that eclectic texts have been chosen as the basis of annotation is that they have been the most readily available electronically. Copyright issues aside, numerous parsed and annotated texts of various editorial texts have been available on the internet for numerous decades.[40]. In recent years digitisation projects have made manuscripts more accessible for study through high quality digital images available online.[41] A few projects have also sought to improve access for non-specialists by providing transcriptions alongside the images.[42] In the case of the Codex Sinaiticus project the transcriptions are also linked word for word to the images going even further to help non-experts to explore the manuscript for themselves.

While manuscript images have become readily available for transcription changes to editorial practices in the digital age have also rendered this step optional for

[39] The OpenText.org annotation is based on the NA^{27} text but since with the exception of the Catholic Epistles the text is the same, in the present article we will use NA^{28} for the mapping as transcriptions of this edition were available in the appropriate format for the collation editor.

[40] For a discussion of copyright issues around the Greek New Testament see Porter, *Linguistic Analysis*, 17-28.

[41] See for example http://www.bl.uk/manuscripts/ http://cudl.lib.cam.ac.uk/ http://www.e-codices.unifr.ch/ http://gallica.bnf.fr/ http://www.mss.vatlib.it/guii/scan/link.jsp http://teca.bmlonline.it/TecaRicerca/index.jsp http://www.digitale-sammlungen.de/ and the New Testament specific websites http://www.csntm.org/Manuscript and http://ntvmr.uni-muenster.de

[42] See for example the Codex Sinaiticus project website hosted by the British Library http://www.codexsinaiticus.org/en/ and the Codex Bezae digital edition available from the university of Cambridge http://cudl.lib.cam.ac.uk/view/MS-NN-00002-00041/194 (select ‘diplomatic transcription’ from the ‘view more options’ menu on the right hand side to see the transcription for the images).

those seeking the electronic text of a manuscript to form the basis of their linguistic annotation. Historically critical editions were created by selecting a base text and recording for each manuscript only those places where it differed from the base text using the resulting tables to create the final text and apparatus for an published edition.[43] Since the late 1990s the editorial teams behind the Editio Critica Maior have adopted a digital editing workflow. This workflow involves making full text transcriptions of each of the manuscripts to be included as witnesses for a particular biblical book.[44] The transcriptions form the basis of computer-assisted collations and genealogical analysis which inform the critical edition but in addition the transcriptions are also made available under a creative commons license.[45] The transcriptions are annotated and stored in TEI P5 XML, the current standard exchange format in the digital humanities.[46]

With annotation there is always a balance to be struck between recording everything one can see in a manuscript, how much is required for the current task and how much time is available for transcribing.[47] The ECM teams record the text of the manuscript in some detail, including marking corrections, capitalisation, some marginalia, nomina sacra and a variety of other abbreviations. Accents (with the exception of places where the accents are required for disambiguation such as future

[43] Parker, "Through a Screen," 396-7 and Parker, *Introduction*, 95-100.

[44] For a more detailed discussion of the development and adoption of the digital workflow see Houghton, "Electronic Scriptorium," 31-37.

[45] Transcriptions made of the Gospel of John by ITSEE/IGNTP can be found here http://www.iohannes.com/XML/start.xml. Transcriptions made by the INTF/IGNTP teams can be found here http://ntvmr.uni-muenster.de/community/vmr/api/transcript/get/?gaNum=01&indexContent=Romans&fullPage=true&format=teiraw (replace the gaNum and the indexContent for other manuscripts and biblical books).

[46] For the general TEI guidelines see http://www.tei-c.org/Guidelines/P5/. For a description of the schema used for the IGNTP transcriptions see Houghton "Electronic Scriptorium".

[47] Houghton, "Electronic Scriptorium," 37-8.

forms of liquid verbs) and script changes (with the exception of supplements) however are not recorded. Some physical information about each manuscript is also recorded, page, column and line breaks are recoded as standard, as well as physical damage to the page, but other features not considered pertinent to the creation of an editorial text such as ink colour, illustrations and chapter numbers or titles are not normally recorded.[48] This shows that just as it is important to be aware of the transmission history of the manuscript you select, if you are using third party transcriptions you must be aware of the original transcription policy. Editorial decisions such as the normalisation of spelling, the addition or removal of punctuation and expansion of abbreviations (e.g. nomina sacra, superline nu and kai compendia) as well as countless other things may have been made in process of transcribing. Even something as simple as how words should be divided is not always straightforward when transcribing Greek manuscripts. As a user of third party transcriptions it is crucial to be aware of these decisions and as a producer of transcriptions it is important that as many of these editorial decisions as possible are documented.[49] These cautions aside, the possibilities created by having full manuscript transcriptions available under a creative commons license are huge and makes the mapping of the OpenText.org annotation to single manuscript texts a realistic possibility.

There are several stages involved in mapping the OpenText.org linguistic annotation from the NA^{27} text to multiple single manuscript texts. The first task is to

[48] The Codex Sinaiticus and Codex Bezae projects mentioned in footnote 42 recorded a far greater amount of details than is typical for transcriptions made by the IGNTP, INTF and ITSEE teams. The transcriptions made as part of the Codex Sinaiticus project could have been used for this study but as their XML schema differs from that currently used and therefore would need a lot of preprocessing before they would work with rest of the workflow. Instead simpler transcriptions available from the INTF were used.

[49] See for example the Introduction to the IGNTP edition of the majuscules of the Gospel of John http://www.iohannes.com/majuscule/index.html

compare the text in the target manuscript to that of the NA^{27}. This will show up places where the text is identical and places where the text differs. The nature of the differences will be varied and the different types of variation will have different levels of impact on the linguistic annotation, understanding the extent and nature of the differences between the texts is essential for planning the annotation mapping. Once the differences have been evaluated a computer assisted mapping process will be used to automate the mapping where there are no changes and flag up the areas where user input is required to check the mapping attempted automatically or, where the changes are too complex, present the text to the user alongside the original linguistic analysis for re-annotation.

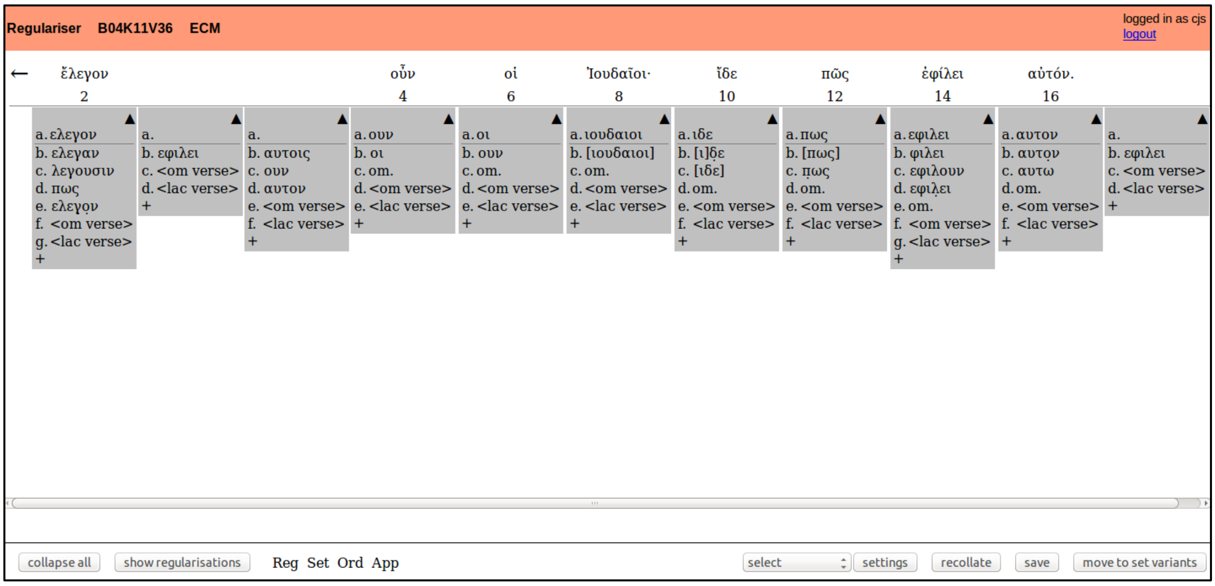
Comparing the Texts

At the same time as making full text manuscript transcriptions available the adoption of a digital editing workflow for the ECM has lead to a focus on the computational collation of texts. There are a few software packages available to choose from[50] but here we will use a collation editor developed by ITSEE at the University of Birmingham. A screenshot of the collation editor in use for the ECM edition of the Gospel of John can be seen in figure 1.[51] The collation editor is a wrapper around the CollateX software[52] and is the

[50] These include CollateX (http://collatex.net/), tustep (http://www.tustep.uni-tuebingen.de/tustep_eng.html) and juxta (http://www.juxtasoftware.org/). The first collation software widely used in the humanities was collate and then collate2 and collate3 created by Peter Robinson, Robinson *Collate*. This software is now obsolete but was certainly the trail blazer in this area and was used by the INTF and IGNTP for making editions of Biblical texts, see Wachtel "Editing the Greek New Testament" and Parker "Electronic Religious Texts". CollateX was designed to be the successor to Robinson's Collate, see http://collatex.net/about/.

[51] The collation editor was developed as part of a collaborative Anglo-German project called 'The Workspace for Collaborative Editing' which was funded by he Arts and Humanities Research Council and the Deutsche Forschungsgemeinschaft between 2010 and 2013.

software used by the ECM teams to make their editions. It is designed to be a fully interactive system allowing editors to go from full text transcriptions of manuscripts to published apparatus. CollateX provides an initial computerised comparison of the texts which the editor can then manipulate in the collation editor interface. It includes tools for regularising unimportant differences such as spelling variation and abbreviations, setting the length of each variant and ordering the readings in each unit. The experience of using the collation editor in the making of the ECM edition of John's gospel is that once regularisation is complete CollateX does a good job of collating the 233 manuscripts in the edition where there are not too many large multi-word omissions, additions or transpositions in a verse. The task of comparing just two texts is of course far simpler than the task of comparing over 200 and therefore it is reasonable to expect that the comparisons produced by CollateX will be a good starting point for this task. This does not mean that the alignment will always be correct but the level of error should be small, and finding and correcting these errors will be far less time consuming than producing the alignments by hand.



ἔλεγον 2 | οὖν 4 | οἱ 6 | Ἰουδαῖοι· 8 | ἴδε 10 | πῶς 12 | ἐφίλει 14 | αὐτόν. 16

a. ελεγον
b. ελεγαν
c. λεγουσιν
d. πως
e. ελεγọν
f. <om verse>
g. <lac verse>
+

a.
b. εφιλει
c. <om verse>
d. <lac verse>
+

a.
b. αυτοις
c. ουν
d. αυτον
e. <om verse>
f. <lac verse>
+

a. ουν
b. οι
c. om.
d. <om verse>
e. <lac verse>
+

a. οι
b. ουν
c. om.
d. <om verse>
e. <lac verse>
+

a. ιουδαιοι
b. [ιουδαιοι]
c. om.
d. <om verse>
e. <lac verse>
+

a. ιδε
b. [ι]δ̣ε
c. [ιδε]
d. om.
e. <om verse>
f. <lac verse>
+

a. πως
b. [πως]
c. π̣ως
d. om.
e. <om verse>
f. <lac verse>
+

a. εφιλει
b. φιλει
c. εφιλουν
d. εφιλ̣ει
e. om.
f. <om verse>
g. <lac verse>
+

a. αυτον
b. αυτọν
c. αυτω
d. om.
e. <om verse>
f. <lac verse>
+

a.
b. εφιλει
c. <om verse>
d. <lac verse>
+

Figure 1: Screenshot of the collation editor

[52] This software was developed by Ronald Dekker and the Hyugens Institute as part of the EU funded interedition project, see www.interedition.eu.

When the collation editor is used for creating an edition of a text the regularisation phase is circular. Regularisation rules are made and then the texts are recollated and over several iterations; as irrelevant differences are removed, the alignment produced by CollateX gradually improves. For just two texts the effort required in this stage is not practical: instead, we want minimal human intervention in the process. In the collation editor a certain amount of preparation happens to all of the data automatically before it is passed to CollateX in order to achieve the most accurate alignment possible from CollateX. All of the text is turned to lower case, markers for unclear and supplied text are removed and simple abbreviations are expanded and superline nus are replaced with the actual character. These changes only affect the data sent to CollateX for alignment; the way the words are displayed in the interface will not necessarily be the same tokens that were sent to CollateX but instead will be determined by the settings selected by the user. This same preparation was used for this research. Alongside this we also used a set of global rules created for the ECM John project that expand all of the common nomina sacra abbreviations to their full text form which is how they appear in NA^{27}. Each book was batch processed verse by verse to produce alignment tables that could be analysed to give an overview of the scale of the challenge ahead. The data was also made available in the collation editor interface so that the results could be more easily compared with the existing annotation to establish impact of the changes.

The extent of the challenge

To assess the scale of the challenge, the first four Pauline epistles were selected for testing. These books were batch processed with the collation editor using NA^{28} as the base text. The collation editor provides an alignment table in JSON[53] for each verse and these alignment tables were analysed for the volume and nature of the differences using a specially written python script. The structure of the tables reveal whether the changes (assuming changes from the base text) are additions, omissions or substitutions, as well as being able to calculate the total number of the verses that differ from NA^{28}. Verses with no differences will be straightforward when it comes to mapping the annotation, those with differences will be more complex and the type of the differences will affect the mapping in different ways. The transcriptions used for this study are those produced by the INTF and are available through the NTVMR.[54]

As well as using Codex Sinaiticus we decided to include another manuscript in the test to see how effectively the methodology used in this study could be generalised to mapping the OpenText.org annotation to other manuscripts. Since Codex Sinaiticus is an example of the Alexandrian text-type which forms the core of our modern critical texts we might expect it to be very similar to the Nestle-Aland text. As a contrast to this we have selected P46. P46 is not fully extant in the four books we have selected: it begins midway through chapter 5 of Romans and has a few lacunose or omitted verses in the other books. This is not of concern for this study as lacunose or omitted verses are generally irrelevant for the annotation mapping process; the only places where this might be of concern is for verses that are partly lacunose or places where an annotation unit wraps verse boundaries and one of the verses involved is lacunose. P46 is an important manuscript in the study of the development of the Pauline Epistles and while

[53] JSON (JavaScript Object Notation) is a standard data exchange format. See http://www.json.org.

[54] http://ntvmr.uni-muenster.de/community/vmr/api/transcript/get/?gaNum=01&indexContent=Romans&fullPage=true&format=teiraw, http://ntvmr.uni-muenster.de/community/vmr/api/transcript/get/?gaNum=P46&indexContent=Romans&fullPage=true&format=teiraw

the manuscript itself is not of particularly great standard it is thought that the *Vorlage* is of particularly high quality.[55] Zuntz points out that '[t]he scribe committed very many blunders'[56] and also notes that 'the omission of whole clauses owing to homoioteleuton [eye-skip] is an outstanding characteristic of P46'[57] Despite this the manuscript has a far lower level of correction than Codex Sinaiticus so in P46 we will ignore all corrections since none in the books we are using have been identified as having been made by the first hand.[58] Zuntz considers P46 to belong to a group of manuscripts representing the proto-Alexandrian text-type which is generally similar to the Alexandrian text-type but which also has many readings more typically found in the Western text-type.[59] That, along with the error prone nature of the text, leads us to predict that P46 will present a greater challenge when it comes to mapping the OpenText.org annotation. It will also serve as a representative of the vast majority of New Testament witnesses which are not of such a high quality as Codex Sinaiticus in order to see how generally applicable the methodology developed might be to a wider range of New Testament texts.

An initial comparison between the number of verses affected by differences in Codex Sinaiticus and P46 is shown in Table 1 and Table 2. This shows that, as expected, P46 has far more verses with differences than Codex Sinaiticus although even in this witness more than half of the verses for each book still have at least one difference from NA^{28}. In P46 this rises to an average of around 70 percent of verses. (In each case percentages are given as percentage of extant verses in the witness for the given book).

[55] Holmes, "Text," 189; Zuntz, *Text*, 56.
[56] Zuntz, *Text*, 18.
[57] Zuntz, *Text*, 19.
[58] If there had been any first hand corrections in the transcriptions we would have replaced the initial first hand text with the first hand correction.
[59] Zuntz considers the proto-Alexandrian text-type to preserve readings which came to be lost from the Alexandrian tradition but remained in the Western tradition. Zuntz, *Text*, 156; Holmes "Text," 199.

There is a remarkable consistency in the percentages between the books in each witness particularly in Codex Sinaiticus. In terms of mapping the annotation this shows that it should be possible to map the OpenText.org annotation directly to Codex Sinaiticus for just under half of all verses although the data for P46 suggests that this figure will be lower for other witnesses. The number of differences per verse will also have an impact on the complexity of the task. Table 3 shows the total number of points of difference between each witness and NA^{28} per verse. OpenText.org has annotation down to the word group level and therefore each word has the potential to require changes to the annotation. For this reason Table 3 is based on the number of words that differ from NA^{28} per verse; percentages are given as the percentage of verses with differences from Tables 1 and 2. This data has not been separated out into individual books as in each witness the distribution curves for the different books are very similar. Table 3 shows that, as well as having fewer verses with differences, in Codex Sinaiticus there are generally fewer differences per verse. Nearly 60 percent of all verses with differences in Codex Sinaiticus have only one word that is different from NA^{28}; the figure

	Extant verses	Verses without changes		Verses with changes	
		Raw count	Percentage	Raw count	Percentage
Romans	432	212	49	220	51
1 Corinthians	437	213	49	224	51
2 Corinthians	256	126	49	130	51
Galatians	149	66	44	83	56

Table 1: The number of verses with and without changes in Codex Sinaiticus when compared with NA^{28}

	Extant verses	Verses without changes		Verses with changes	
		Raw count	Percentage	Raw count	Percentage
Romans	247	82	33	165	67
1 Corinthians	433	125	29	308	71
2 Corinthians	254	66	26	188	74
Galatians	140	35	25	105	75

Table 2: The number of verses with and without changes in P46 when compared with NA^{28}

for P46 is roughly half that. Nearly a quarter of all the verses with differences in P46 involve five or more words but that figure is dramatically lower in Codex Sinaiticus at around three percent. A closer examination of the verses with a difference of 10 or more words shows that all but two of them are caused by the presence of a lacuna (typically at the bottom of pages where the leaves of P46 are damaged). As mentioned above we know from work on the ECM of John that CollateX does not always do a good job of alignment when large lacuna are involved and therefore misalignment of the text that is present could be inflating these counts above the number of words actually missing. The two examples not influenced by lacuna are 2Cor 1:7 and 2Cor 8:19. These two cases are examples of the homoioteleuton noted by Zuntz and discussed above. Based on Zuntz's observation we should expect that the instance of omissions in P46 will be high. If these homoioteleuton are of clauses as hinted at by Zuntz and CollateX has managed to correctly align the extant text then their impact on the OpenText.org annotation should be minimal since most of the annotation works on the clause level or below. This demonstrates that the type of changes as well as the extent of them will have an impact on how complex the annotation mapping will be.

Total differences in words per verse	01		P46	
	Raw count	Percentage	Raw count	Percentage
1	388	59	240	31
2	168	26	179	23
3	62	9	114	15
4	22	3	52	7
5	11	2	40	5
6	3	0.5	26	3
7	2	0.3	20	3
8	0	0	20	3
9	1	0.2	17	2
10+	0	0	57	7

Table 3: The number of differences (counted in words) per verse for each witness against NA^{28}

The CollateX alignment tables can be easily analysed to see into which of three categories the differences can be placed. The three categories are addition, omission and substitution. The collation output has not been checked for accuracy so it is not possible to say for certain that a difference should definitely belong to a particular category just that it is the category in which the automated collation has assigned it.

	01		P46	
	Raw count	Percentage	Raw count	Percentage
Addition	78	7	143	5
Omission	115	11	1597	58
Substitution	883	82	1025	37

Table 4: The type of changes (counted in words) for each witness against NA^{28}

Even so an examination of the categories should give a good overview of the type of changes present in the witnesses. Table 4 shows the volume by word of each type of correction in the two witnesses. The contrast here is stark. Over 80 percent of the differences in Codex Sinaiticus are categorised as substitutions with omissions and additions playing a far lesser role. P46 meanwhile has a far higher percentage of omissions as we might expect. While a good number of these will probably be caused by the more fragmentary nature of the manuscript they will also include the omissions due to homoioteleuton noted by Zuntz. In P46 there are also still a large number of substitutions but again additions play a far more minor part.

The importance of the substitution category in both manuscripts but particularly in Codex Sinaiticus is good news for the mapping procedure. In most cases substitutions will be easier to handle than omissions or additions (with the exception of potential whole clause omissions noted earlier). A quick scan through the list of substitutions for each manuscript shows that many of them are merely orthographic differences such as the change of ει to ι (e.g. ημις for ημεις, γεινεσθε for γινεσθε), or the reduction or doubling of consonants (e.g. απολυε for απολλυε, ανηγγελλη for ανηγγελη) which will make no difference to the annotation. Some are also caused by the nomina sacra that have not been encountered in our collation of John and are therefore not yet on the list of nomina sacra expansions that we have applied (e.g. εσταν for εσταυρωσαν). A closer analysis of the data for Romans in Codex Sinaiticus show that of the 342 points of difference 237 belong to one of these orthographic categories. This is nearly 70 percent of all differences. Once the substitutions that can be classified as orthographic are removed from the data only 74 verses in Romans differ from NA^{28} in Codex Sinaiticus, this is around 17 percent of the extant verses, a considerable reduction. It is possible that rules could be generated that would allow these purely orthographic changes to be detected so that verses containing only such differences could be treated as identical verses and have the annotation mapped automatically.

The remaining differences are the those which will, or which at least could, require a change in annotation. Of the remaining examples some are differences in the grammatical form of the word which may or may not require a change in annotation above the word level (e.g. οφιλοντες for οφειλετε, σφραγισαμενοις for σφραγισαμενος). Others could be orthographic but could also be grammatical changes depending on the context (e.g. επιμενομεν for επιμενωμεν, διωκομεν for διωκωμεν). Others are lexical substitutions which make sense in the context (e.g. κακον for φαυλον, πετρος for κηφας, οτι for και). Others are difficult to see as genuine substitutions and are most likely due to incorrect alignment (e.g. Romans 9:16 where a word order change in P46 means τρεχοντος has been classified as a substitute for θελοντος and vice versa and Romans 15:17 in P46 where different wording has resulted in ην being treated as a substitute for ουν and εχω as a substitute for την. In this particular case there is an εχω in NA^{28} but the change in word order has lead to this being classified as an omission in P46).

This survey of the extent and nature of the changes in Codex Sinaiticus and P46 has shown that the level of difference is fairly extensive, even in the manuscript which is one of the closest to the critical text currently used for the OpenText.org annotation. However if the orthographic changes could be identified automatically then the problem would be reduced considerably. The disparity in both volume and type of differences between the two manuscripts selected is also large and therefore suggests that mapping the annotation to multiple manuscripts may not be a practical option as a solution that works well for one manuscript may not work as well for another. The remainder of this paper then will focus on Codex Sinaiticus, Porter's suggestion for a replacement text for NA^{28} and the text which was initially intended to form the annotation base for OpenText.org. We cannot guarantee that the alignment provided by CollateX is always accurate and this might cause further problems for the mapping process. We can however guarantee that all the verses marked by the collation editor as not having any changes are genuinely the same and can have the annotation transferred automatically. Most importantly this survey has provided a list of verses which have differences from our current text and which therefore need further investigation to see what impact the differences in them might have on our annotation.

The linguistic impact of differences

The current OpenText.org syntactically annotated Greek New Testament separates this analysis into two main levels: the word group level and the clause level. The word group level focuses on the direct connections between pairs of words. It is based on dependency grammar where pairs of related words have asymmetric relations so one is the head and the other the modifier. A modifying word can belong to one of four categories: specifier (sp) where the modifier provides specification of the modified word (e.g. articles and prepositions); definer (df) where the modifier provides further definition of the modified word (e.g. apposition, attributive adjectives, predicate adjectives); qualifier (ql) where the modifier limits the scope of the modified word (e.g. genitive and dative modifiers and word group negation); and relator (rl) where the modifier is in a prepositional relationship with the modified word (e.g. prepositional phrases within word groups). Each word in a word group can be modified by other words in the word group and therefore these relations are often nested.[60] The word group relations are visualised in the OpenText.org annotation using a series of nested boxes as in the examples below.[61] The modifier slots are labelled in the shaded row and words are placed in the column that defines the way they modify the numbered word in the top row. The first column labelled cn is used for conjunctions. Conjunctions are only included within word groups if they function at word group level (e.g. to join two modifiers). Conjunctions which function to join word groups are marked at clause level.[62] The clause level annotation groups words together into functional clause components. The four core clause components are: subject (S), which is the grammatical subject of the clause; predicator (P), which is the verbal element of the clause; complement (C), which is the word or word groups that complete the predicate of the clause; adjunct (A), which are adverbs, adverbial clauses and prepositional phrases that modify the verb. In addition to these core components OpenText.org also uses two further categories: addressee (add), which is used for vocative forms or other forms functioning to call attention; conjunction (cj), which is for words that function to link clauses together.[63] There are three levels of clause in the OpenText.org model:

[60] O'Donnell, *Corpus Linguistics*, 179-80.
[61] As proposed in O'Donnell, *Corpus Linguistics*, 180.
[62] O'Donnell, *Corpus Linguistics*, 175.
[63] O'Donnell, *Introducing*.

a) primary clauses are independent clauses which usually contain a finite verb form and are not subordinate to any other clause;

b) secondary clauses are clauses that are subordinate to another clause. These are commonly relative clauses and clauses beginning with words such as ὡς/καθώς and ὅτε/ὅταν although non-embedded participial and infinitive clauses are also classed as secondary;

c) embedded clauses are clauses that occur inside a component of another clause. These clauses usually have non-finite verb forms but finite clauses can also be embedded.[64]

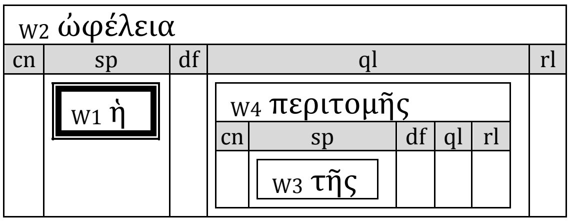
In the following sections we explore the impact of switching from an eclectic text to a single manuscript on higher levels of linguistic annotation such as these by focussing on the differences between the texts of NA^{28} and Codex Sinaiticus in Romans. The data is organised into the following categories: 1. omissions, 2. insertions, 3. wording changes and 4. word order changes.[65]

Omissions

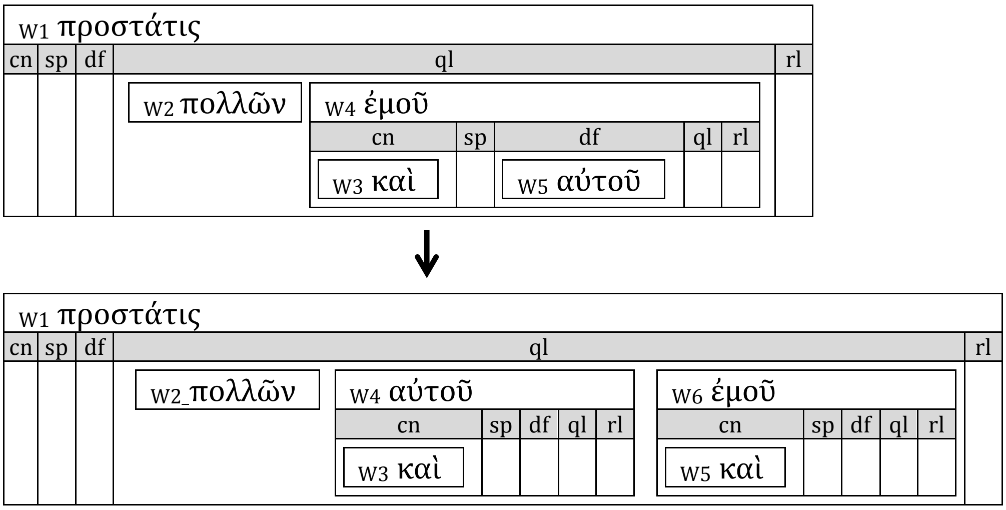
The impact of omissions on the annotation varies hugely, according to the nature of the omission. The omission of the article, which happens relatively frequently in our example text, requires minimal changes to the word group annotation. In this example from Rom 3:1 Codex Sinaiticus omits the article ἡ from the phrase ἡ ὠφέλεια τῆς περιτομῆς.

[64] O'Donnell, *Introducing*.

[65] In this section we will refer to these categories as changes to, omissions from and insertions into the NA^{28} by Codex Sinaiticus purely because NA^{28} is the text currently used for the OpenText.org annotation and it is therefore changes to this text that we are interested in. It is not intended to reflect the reality of text transmission.



This change would only require removing the specifier-specified relation between ἡ and ὠφέλεια; all other relations in the word group would remain the same. The same is true of the other examples of article omission such as Rom 3:12 ούκ ἔστιν [ὁ] ποιῶν χρηστότητα; 3:25 ἱλαστήριον διὰ [τῆς] πίστεως; 4:11 καὶ αὐτοῖς [τὴν] δικαιοσύνην and 10:5 τὴν δικαιοσύνην τὴν έκ [τοῦ] νόμου. There are no straightforward examples of omitted head terms in the Codex Sinaiticus text of Romans but in Rom 16:2 we do have an example of insertion which involves a rearrangement of the word group resulting in the loss of a head term. In this example Codex Sinaiticus replaces the last three words of the phrase προστάτις πολλῶν έγενήθη καὶ έμοῦ αὐτοῦ in NA^{28} with καὶ αὐτοῦ καὶ έμοῦ.[66] In NA^{28} αὐτοῦ is a definer of the head term έμοῦ but the addition of the conjunction makes this phrase two word groups rather than one and the first word group loses its head term as έμοῦ becomes the head term of the second word group.



[66] έγενήθη is not part of the word group here but instead the predicate of the main clauses which is embedded within this complement word group.

With only a single modifier of ἐμοῦ the decision about which to promote is simple and could perhaps be automated, but in more complex situations annotator input will be required to deal with this.

The examples of omissions discussed so far have no impact on the clause level annotation. The omission of a modifier within a word group, as in our first example, will have no affect on the clause component in which it is contained. Similarly, if a clause component contains more than one word group and one of them is removed, or rearranged as in our second example, this will likely not change the category of the clause component itself. The simplest of the omissions that do impact the clause annotation are examples where the words omitted comprise a complete clause component that can be removed from the clause structure. For example in Rom 11:21 the first two words, μή and πως, are omitted in Codex Sinaiticus. This requires removal of the first two adjuncts in the clause.

Primary	A	A	A	C	P
	μή	πως	οὔδὲ	σοῦ	φείσεται

A similar example can be seen in Rom 16:7 where Codex Sinaiticus omits the nominative plural relative pronoun οἳ in the second of a pair of secondary clauses, which has a S-cj-A-P-A structure. This results in the removal of the subject component from the second secondary clause.

Secondary 1	S	P	C
	οἵτινές	εἴσιν	ἐπίσημοι ἐν τοῖς ἀποστόλοις

Secondary 2	S	cj	A	P	A
	οἳ	καὶ	πρὸ ἐμοῦ	γέγοναν	ἐν Χριστῷ

There are no examples of longer omissions (more than a couple of words) in the text of Romans in Codex Sinaiticus which have not been identified and corrected by a scriptorium hand. There are two examples in 1 Corinthians, although even in these cases the omitted text has been added in the lower margin of the page by one of the later correctors (C^{a}). In 1Cor 15:27 the first clause is omitted, but as the clause diagrams show this is a primary clause with no dependents so the full clause can be omitted without requiring any other changes to the annotation other than to connect the second primary clause in the diagram (labelled Primary 2) to the primary clause immediately before the one deleted.

Primary 1	C	cj	P	A
	πάντα	γὰρ	ὑπέταξεν	ὑπὸ τοὺς πόδας αὐτοῦ

Secondary 1 (connect P2)	cj	cj	P
	ὅταν	δὲ	εἴπῃ

Secondary 2 (connect S1)	cj	S	P
	ὅτι	πάντα	ὑποτέτακται

Primary 2 (connect P1)	C
	δῆλον

<table>
<tr><td rowspan="2">Secondary 3
(connect P2)</td><td>cj</td><td>A</td></tr>
<tr><td>ὅτι</td><td><table><tr><td rowspan="2">Embedded 1</td><td>P</td><td>C</td><td>C</td></tr><tr><td>ἐκτὸς τοῦ ὑποτάξαντος</td><td>αὐτῷ</td><td>τὰ πάντα</td></tr></table></td></tr>
</table>

In 1Cor 15:54 there is an example of what is most likely homoioteleuton in which the words φθαρτὸν τοῦτο ἐνδύσηται ἀφθαρσίαν καὶ are omitted from Codex Sinaiticus. This results in two parallel secondary clauses being combined into one in the clause annotation. The following two clauses in NA^{28},

Secondary 1	cj	cj	S	P	C
	ὅταν	δὲ	τὸ φθαρτὸν τοῦτο	ἐνδύσηται	ἀφθαρσίαν

Secondary 2 (connect S1)	cj	S	P	C
	καὶ	τὸ θνητὸν τοῦτο	ἐνδύσηται	ἀθανασίαν

become this single clause in Codex Sinaiticus:

Secondary 1	cj	cj	S	P	
	ὅταν	δὲ	τὸ θνητὸν τοῦτο	ἐνδύσηται	τὴν ἀφθαρσίαν

Insertions

Insertions will always necessitate additional annotation since the new word or words must be assigned to word groups and clauses. The simplest cases are additions to the existing word group structures, as these do not necessitate changes or additions to the clause level annotation. In Rom 3:5, for example, the phrase ὁ θεὸς ὁ ἐπιφέρων τὴν ὀργήν has αὐτοῦ added at the end in Codex Sinaiticus and only necessitates the addition of a qualifier relation between αὐτοῦ and ὀργήν in an existing word group.

w2 ὀργήν				
cn	sp	df	ql	rl
	w1 τὴν			

→

w2 ὀργήν				
cn	sp	df	ql	rl
	w1 τὴν		w3 αὐτοῦ	

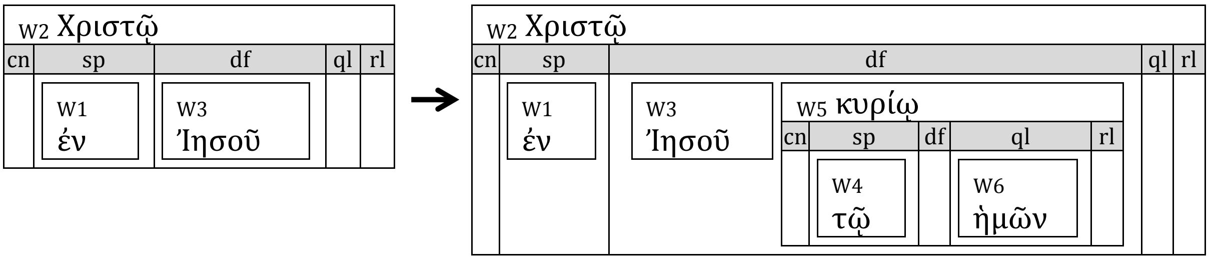
Similarly in Rom 5:18 Codex Sinaiticus adds ἀνθρώπου as a qualifier of παραπτώματος in the phrase Ἄρα οὖν ὡς δι’ ἑνὸς ἀνθρώπου παραπτώματος εἰς πάντας ἀνθρώπους εἰς κατάκριμα, leading to the following change in word group structure.

w4 παραπτώματος				
cn	sp	df	ql	rl
	w1 ὡς w2 δι’	w3 ἑνὸς		

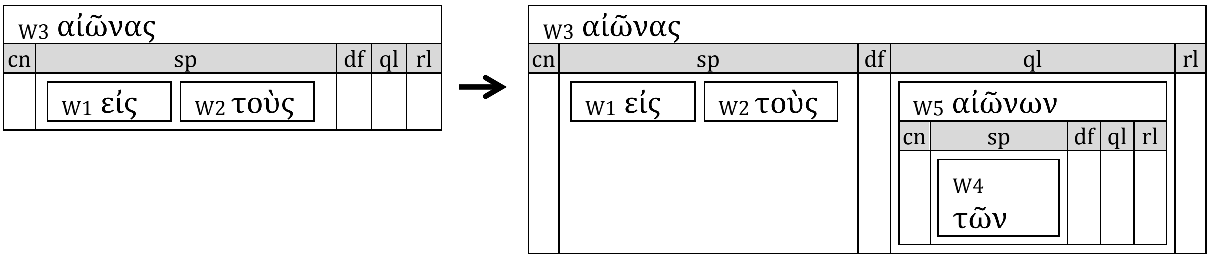
→

w5 παραπτώματος				
cn	sp	df	ql	rl
	w1 ὡς w2 δι’	w3 ἑνὸς	w4 ἀνθρώπου	

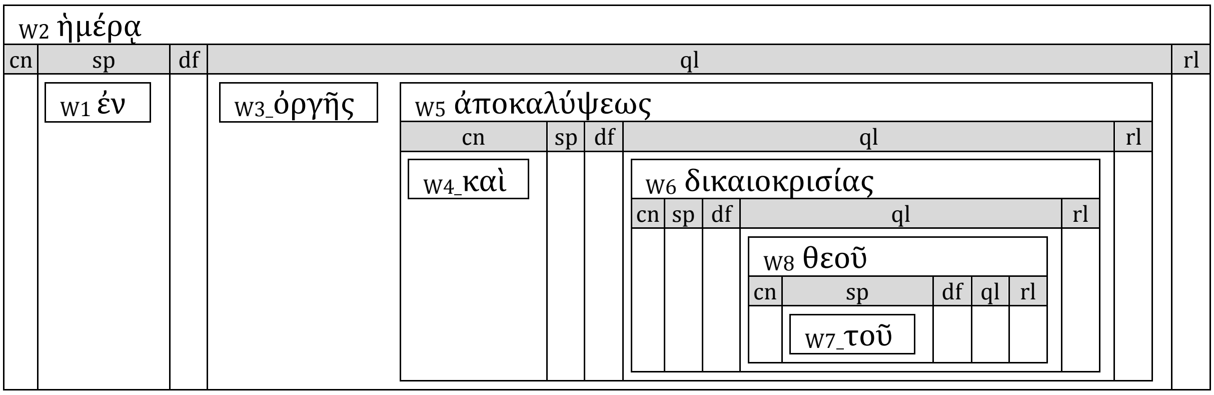
Even when the additions are of multiple words if they are only additions to the word group structure the impact is minimal. For example in 6:11 the word group ἐν Χριστῷ Ἰησοῦ has the addition of the defining phrase τῷ κυρίῳ ἡμῶν immediately following it. This insertion requires a change in the word group annotation but again does not impact any of the existing word group relations or the annotation at the clause level.



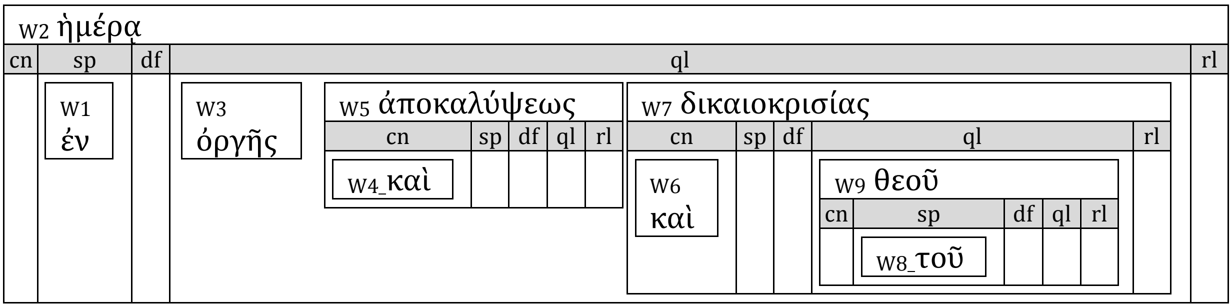
In the final clause of Romans, at 16:27, Codex Sinaiticus adds the words τῶν αἰῶνων to the word group εἰς τοὺς αἰῶνας, resulting in a qualifier-qualified dependency between αἰῶνων and αἰῶνας. Again no changes are needed in the clause annotation.



In Rom 2:5 the addition of a single conjunction necessitates a reorganisation of the existing word group. Again the changes required are confined to the word group level. The text in question here is ἐν ἡμέρᾳ ὀργῆς καὶ ἀποκαλύψεως καὶ δικαιοκρισίας τοῦ θεοῦ with the underlined καὶ being the insertion in Codex Sinaiticus. In the NA^{28} word group annotation, there is a qualified-qualifier relation between ἀποκαλύψεως and δικαιοκρισίας, as shown below.



The insertion of the conjunction in Codex Sinaiticus places δικαιοκρισίας τοῦ θεοῦ on equal level with όργῆς and άποκαλύψεως and thus δικαιοκρισίας becomes an additional qualifier of ἡμέρᾳ and καὶ is added as a connective relation of δικαιοκρισίας resulting in the following word group structure.



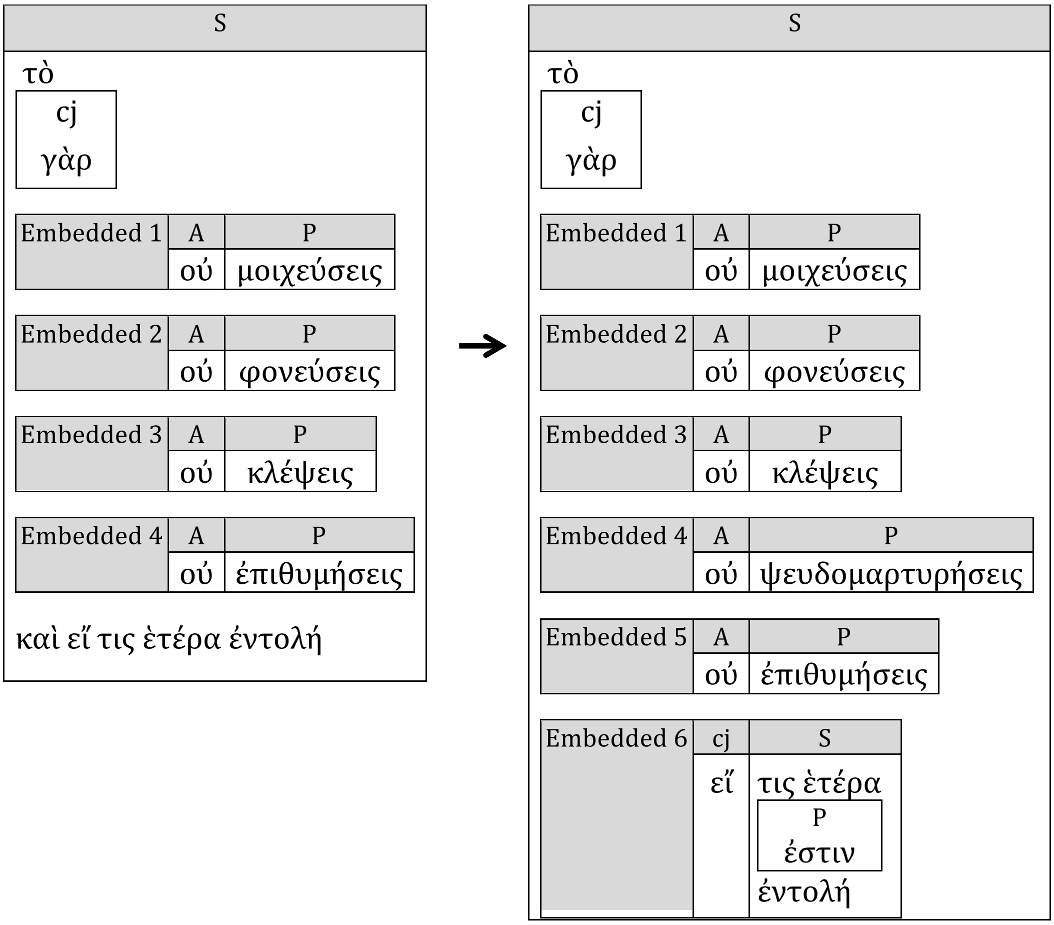
Other changes require addition components to be added to clauses at the clause level (in addition to creating new word groups at the word group level). In Rom 8:34 for example Codex Sinaiticus adds έκ νεκρῶν at the end of the phrase Χριστὸς Ἰησοῦς ὁ άποθανών, μᾶλλον δὲ έγερθείς. In the current clause level annotation this phrase in its entirety is the complement of a clause and includes two embedded clauses. The addition in Codex Sinaiticus requires an adjunct to be added to the second of the two embedded clauses. The resulting clause structure for the final embedded clause is shown below.

Embedded	A	cj	P	A
	μᾶλλον	δὲ	έγερθείς	έκ νεκρῶν

Some additions require new clauses to be created for example in Rom 11:2 (ούκ άπώσατο ὁ θεὸς τὸν λαὸν αύτοῦ ὃν προέγνω. ἢ ούκ οἴδατε έν Ἠλίᾳ τί λέγει ἡ γραφή, ὡς έντυγχάνει τῷ θεῷ κατὰ τοῦ Ἰσραήλ;) Codex Sinaiticus appends λέγων to the end of the verse before the quotation. This requires an additional adjunct in the final clause of this verse with the adjunct being comprised of an embedded clause of which λέγων forms the lone predicate component.

<table>
<tr><th>Secondary</th><th>cj</th><th>P</th><th>C</th><th>A</th><th>A</th></tr>
<tr><td></td><td>ὡς</td><td>έντυγχάνει</td><td>τῷ θεῷ</td><td>κατὰ τοῦ Ἰσραήλ</td><td><table><tr><th>Embedded</th><th>P</th></tr><tr><td></td><td>λέγων</td></tr></table></td></tr>
</table>

A more substantial example of this can be found in Rom 13:9. There are two additions here which affect the subject component of the single clause that represents the whole verse. Codex Sinaiticus adds the extra commandment ού ψευδομαρτυρήσεις into the list of commandments and also the verb έστιν later on. The full phrase in Codex Sinaiticus (with additions underlined) is τὸ γὰρ ού μοιχεύσεις, ού φονεύσεις, ού κλέψεις, ού ψευδομαρτυρήσεις, ούκ έπιθυμήσεις, καὶ εἴ τις ἐτέρα έστιν έντολή. These two additions require two additional embedded clauses to be added to the already complex subject component.[67]



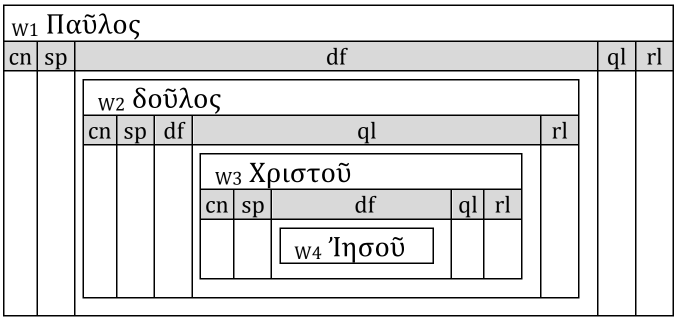
[67] This diagram only shows the subject component of the main clause. The full clause diagram can be seen here http://www.opentext.org/texts/NT/Rom/view/clause-ch13.v0.html

Wording Changes

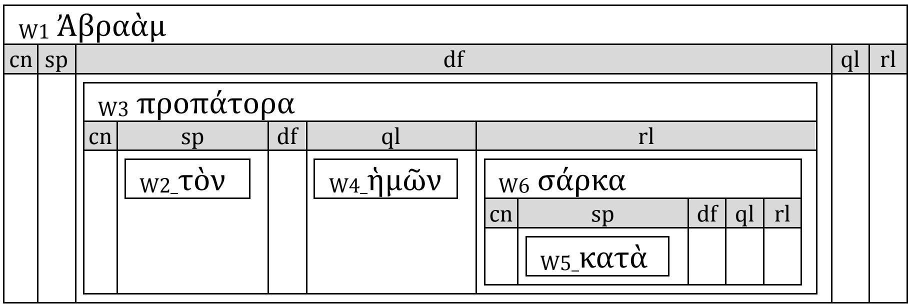
Some of the observed changes between the NA^{28} text and Codex Sinaiticus involve individual word changes that do not require the word group or clause level annotation to be altered in any way, with the orthographic changes noted above included this would be by far the largest category of changes in Codex Sinaiticus. These are often lexical changes such as in Rom 10:17 where διὰ ῥήματος Χριστοῦ in NA^{28} becomes διὰ ῥήματος θεοῦ in Codex Sinaiticus. Here the qualifier-qualified relation between ῥήματος and Χριστοῦ is the same for ῥήματος and θεοῦ, so no changes to the word group annotation are required beyond changing the word itself. Similarly, in Rom 8:35 in the word group ἀπὸ τῆς ἀγάπης τοῦ Χριστοῦ, Codex Sinaiticus replaces the Χριστοῦ in NA^{28} with θεοῦ as a qualifier of ἀγάπης. The first clause of Rom 14:20 has an example of a simple lexical change in the predicator with κατάλυε in NA^{28} replaced by ἀπόλλυε in Codex Sinaiticus. This requires no changes to either level of annotation. Likewise, the last clause of Rom 9:26 τὸ ὑπόλειμμα σωθήσεται in NA^{28} becomes τὸ κατάλειμμα σωθήσεται in Codex Sinaiticus. The S-P clause structure and specifier-specified dependency relations in the first word group are not affected by this purely lexical change. In Rom 3:19 there is another switch of verb from λέγει in NA^{28} to λαλεῖ in Codex Sinaiticus which has no impact on any of the linguistic annotation. Where wording changes are combined with related insertions the insertion means that annotation changes will be required. There is an example of this in Rom 15:32 where διὰ θελήματος θεοῦ in NA^{28} becomes διὰ θελήματος Ἰησοῦ Χριστοῦ in Codex Sinaiticus. This introduces no changes to the clause level annotation but at the word group level, Ἰησοῦ qualifies θελήματος in place of θεοῦ and Χριστοῦ is added as a definer of Ἰησοῦ.

Word Order Changes

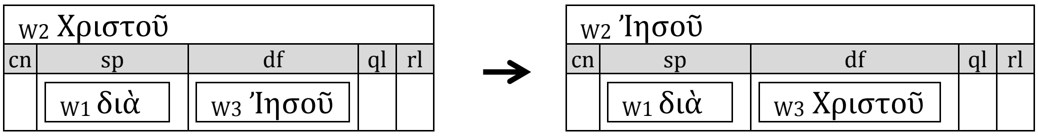
Word order changes can have an impact on the OpenText.org annotation at both clause and word group levels. Rom 1:1 is an example of a straightforward word order change that has minimal impact. NA^{28} reads Παῦλος δοῦλος Χριστοῦ Ἰησοῦ whereas Codex Sinaiticus reads Παῦλος δοῦλος Ἰησοῦ Χριστοῦ. The word group annotation for the NA^{28} text is below.



The word pair Χριστοῦ Ἰησοῦ is an example of two co-defining words that might be seen in equal relation to other. The definer relation captures this grammatical situation and in the NA^{28} annotation Χριστοῦ is the head word modified by Ἰησοῦ to follow word order. The change in word order in Codex Sinaiticus results in a shift of these positions with Ἰησοῦ becoming the head and Χριστοῦ the defining modifier. Not all of the relations at word group level would be affected in this way by a change of word order. For example, if δοῦλος Χριστοῦ Ἰησοῦ became Χριστοῦ Ἰησοῦ δοῦλος the qualifier relation would remain between δοῦλος and Χριστοῦ. This would be the same for words with a relator relation. Consider the word group Ἀβραὰμ τὸν προπάτορα ἡμῶν κατὰ σάρκα in Rom. 4.1 where there is such a relation between προπάτορα and σάρκα.



Imagine a word order change that placed κατὰ σάρκα before προπάτορα, that is, τὸν κατὰ σάρκα προπάτορα ἡμῶν. In this case the roles would be unchanged. The fourth word group relation, the specifier-specified role, such as that between τὸν and προπάτορα, is unlikely to be found in alternate word orders for grammatical reasons. That is, it is not possible to have προπάτορα τὸν and have the relation remain. Instead the relation between the two words would need to be removed, as the article would no longer be acting as a specifier of προπάτορα. It therefore seems possible to derive mapping rules for the word group annotation changes needed to capture word order changes between the NA^{28} text and a single manuscript. A slightly more complex version of the word order change in Rom 1:1 can be found in Rom 2:15. Here again we have an inversion of Χριστοῦ and Ἰησοῦ, but as part of the prepositional phrase διὰ Ἰησοῦ Χριστοῦ. In this instance two 'rewiring' changes are required: 1. the specifier-specified relation between διὰ and Χριστοῦ is now between διὰ and Ἰησοῦ, and 2. the roles in the definer-defined pair Ἰησοῦ Χριστοῦ are switched.



The first clause of Rom. 15.2 is an example of a word order change that impacts the clause level annotation. The NA^{28} text reads καὶ πάλιν Ἠσαΐας λέγει and Codex Sinaiticus reads καὶ πάλιν λέγει Ἠσαΐας. The annotated clause structure is cj-A-S-P, so this change results in a cj-A-P-S structure, with the position of the predicator and subject components.

cj	A	S	P
καὶ	πάλιν	Ἠσαΐας	λέγει

→

cj	A	P	S
καὶ	πάλιν	λέγει	Ἠσαΐας

A similar example that results in a reordering of clause components is found in an embedded clause in Rom 15:32 where ἵνα ἐν χαρᾷ ἐλθὼν with a cj-A-P structure in NA^{28} is changed to ἵνα ἐλθὼν ἐν χαρᾷ, a cj-P-A structure, in Codex Sinaiticus. Word order changes such as these are most likely to have implications for the annotation when the words involved are head terms in their respective word groups. Therefore, these cases can be identified and changes attempted automatically, and then flagged for annotator verification.

More complex changes can also be required because of word order changes. An example of this can be seen in Rom 10:5. Here the word ὅτι is moved in Codex Sinaiticus to a position just a few words earlier in the verse but the resulting syntactic changes require the reconfiguration of two clauses. The clause annotation for these two clauses in NA^{28} is shown below.[68]

Primary 1	S	cj	P	A
	Μωϋσῆς	γὰρ	γράφει	τὴν δικαιοσύνην τὴν ἐκ τοῦ νόμου

<table><tr><td rowspan="2">Primary 2</td><td>cj</td><td>S</td><td>P</td><td>A</td></tr><tr><td>ὅτι</td><td><table><tr><td rowspan="2">Embedded 1</td><td>P</td><td>C</td></tr><tr><td>ὁ ποιήσας</td><td>αὐτὰ</td></tr></table>ἄνθρωπος</td><td>ζήσεται</td><td>ἐν αὐτοῖς</td></tr></table>

In Codex Sinaiticus the word order is changed, with ὅτι being placed after the predicate of the first primary clause changing the clause boundaries. Alongside this, perhaps even because of this, there are also changes to the structure of the second clause. The αὐτὰ in

[68] In this example the embedded clause in the subject of the second primary clause is a definer of ἄνθρωπος at the word group level.

the embedded clause is omitted: its place is taken by τὴν δικαιοσύνην τὴν ἐκ νόμου from the first primary clause. The final αὐτοῖς is also replaced by αὐτῇ, changing the antecedent from αὐτὰ to τὴν δικαιοσύνην. The clause annotation of this verse in Codex Sinaiticus is shown below.

Primary 1	S	cj	P
	Μωϋσῆς	γὰρ	γράφει

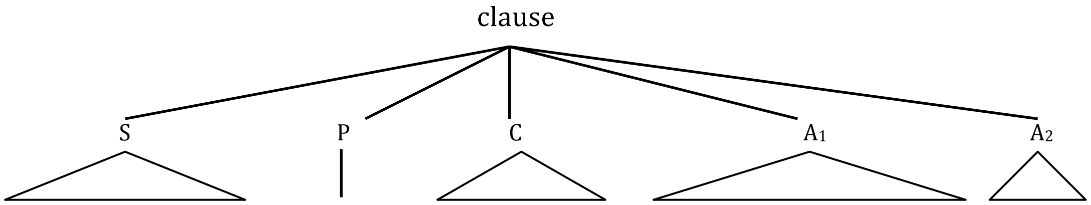
<table><tr><td rowspan="2">Primary 2</td><td>cj</td><td>S</td><td>P</td><td>A</td></tr><tr><td>ὅτι</td><td><table><tr><td rowspan="2">Embedded 1</td><td>C</td><td>P</td></tr><tr><td>τὴν δικαιοσύνην τὴν ἐκ νόμου</td><td>ὁ ποιήσας</td></tr></table>ἄνθρωπος</td><td>ζήσεται</td><td>ἐν αὐτῇ</td></tr></table>

This level of change is not possible to automate completely. However, since word order changes are typically signalled by the collation editor through pairs of omissions and additions of the same word within the same verse, it should be possible to identify places where this is happening and present them to an annotator for evaluation.

Mapping the annotation

In the existing OpenText.org annotation the clause and word group levels have been treated as related but separate streams of annotation. Part of the reason for this decision was that they capture different type of grammatical analysis. The clause annotation is a constituency analysis that uses a relatively flat hierarchy to capture the functional components in each clause. For example, in this clause from Rom. 16.20, there are six components:

||S ὁ (cj δὲ) θεὸς τῆς εἰρήνης |P συντρίψει |C τὸν Σατανᾶν |A ὑπὸ τοὺς πόδας ὑμῶν |A ἐν τάχει ||

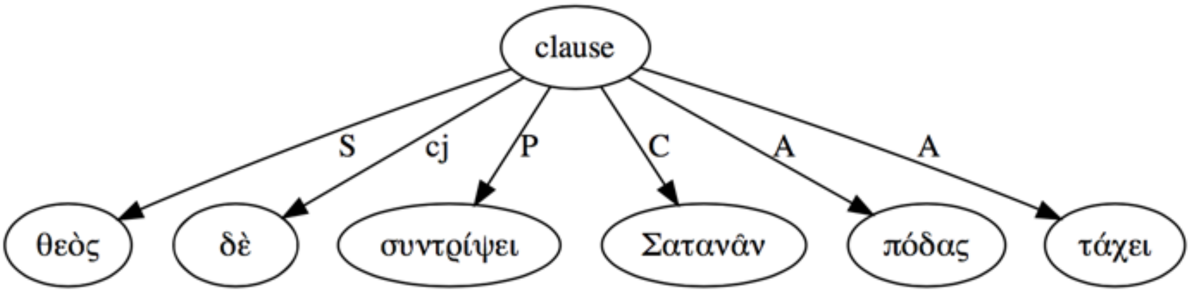


	S	P	C	A	A
	ὁ cj δὲ θεὸς τῆς εἰρήνης	συντρίψει	τὸν Σατανᾶν	ὑπὸ τοὺς πόδας ὑμῶν	ἐν τάχει

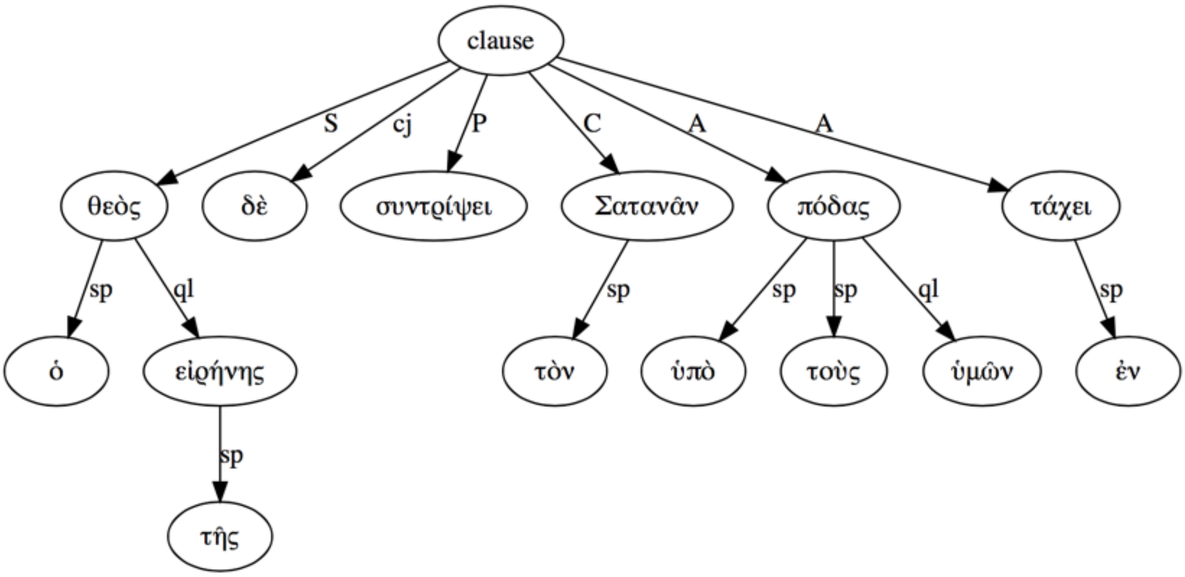
The word group level annotation, in contrast, is a dependency based analysis that captures the relations between pairs of words. For example, in the subject of this clause there is a single word group of which θεὸς is the head. This head term is modified in turn by ὁ, through a specifier relation, and εἰρήνης, through a qualifier relation. Finally, the qualifier εἰρήνης itself has a specifier in the form of the article τῆς.

<table>
<tr><td colspan="5">w_3 θεὸς</td></tr>
<tr><td>cn</td><td>sp</td><td>df</td><td>ql</td><td>rl</td></tr>
<tr><td></td><td>w_1 ὁ</td><td></td><td><table>
<tr><td colspan="5">w_5 εἰρήνης</td></tr>
<tr><td>cn</td><td>sp</td><td>df</td><td>ql</td><td>rl</td></tr>
<tr><td></td><td>w_4 τῆς</td><td></td><td></td><td></td></tr>
</table></td><td></td></tr>
</table>

These two levels of annotation could be integrated into a single dependency based analysis. Under such an analysis the clause would serve as the root node that would have links to the head terms of the word groups within the clause components:



Then the remaining words are linked to their head terms and the relation between them used as the label on the edge:



The same analysis is captured, but the second example offers a single dependency graph. This representation has particular advantages for adapting the annotation to different base texts. Further, it would open the possible of capturing different annotations over different base texts in a single graph.

This evaluation of the scale of the challenge suggests that mapping the OpenText.org annotation to single manuscript witness is an achievable goal. It has also highlighted that the nature and scale of the challenge will differ depending on the manuscript selected. The collation editor seems to do a good job of highlighting the nature of the differences that exist between two texts. Algorithms could also be established for working out which differences are most likely to be orthographic, and can therefore be ignored, and which will require closer inspection. Some elements of the annotation mapping processes could also be automated as detailed above. This study has also highlighted the need to look again at the XML structure used to store the annotation with a view to moving towards a more flexible solution. In future work, we plan to explore the use of a full dependency style representation of the OpenText.org syntactical annotation with multiple base texts to put the choice of text in the users' hands.

Bibliography

Holmes, Michael W. "The Text of P46: Evidence of the Earliest 'Commentary' on Romans?" in *New Testament Manuscripts: Their Texts and Their World* edited by Thomas J. Kraus & Tobias Nicklas, TENT 2. Leiden: Brill, 2006.

Houghton, Hugh A. G. "The Electronic Scriptorium: Markup for New Testament Manuscripts" in *Digital Humanities in Biblical, Early Jewish and Early Christian Studies*, edited by Claire Clivaz, Andrew Gregory and David Hamidovic, 31-60. Leiden: Brill, 2013.

Milne, Herbert J. M. and Theordore C. Skeat. *Scribes and Correctors of the Codex Sinaiticus*. London: British Museum, 1938.

O'Donnell, Matthew Brook. *Corpus Linguistics and the Greek of the New Testament*. New Testament Monographs 6, Sheffield: Sheffield Phoenix Press, 2005.

O'Donnell, Matthew Brook. *Introducing the OpenText.org Syntactically Analyzed Greek New Testament*. No pages, Online http://www.opentext.org/resources/articles/a8.html

Parker, David C. "Through a Screen Darkly: Digital texts and the New Testament" *Journal for the Study of the New Testament* 25 (2003) 395-411.

Parker, David C. "Electronic Religious Texts: The Gospel of John" in *Electronic Text Editing* edited by Lou Burnard, Katherine O'Brian O'Keeffe and John Unsworth, 202-205. New York MLA, 2006.

Parker, David C. *An Introduction to the New Testament Manuscripts and Their Texts*. Cambridge: Cambridge University Press, 2008.

Petersen, William L. "What Text Can New Testament Textual Criticism Ultimately Reach?" in *New Testament Textual Criticism, Exegesis, and Early Church History: A Discussion of Methods*, CBET 7, Edited by Barbara Aland and Joël Delobel, 136-52. Kampen: Kok Pharos, 1994.

Porter, Stanley E. *How We Got the New Testament: Text, Transmission, Translation*. Grand Rapids, MI: Baker Academic, 2013.

Porter, Stanley E. *Using Codex Sinaiticus as an Alterative to a Modern Eclectic Text*. No pages, Online http://www.opentext.org/resources/articles/a1.html

Porter, Stanley E. "Why so Many Holes in the Papyrological Evidence for the Greek New Testament?" in *The Bible as Book: The Transmission of the Greek Text*, edited by Scot McKendrick and Orlaith A. O'Sullivan, 167-186. London: The British Library, 2003.

Porter, Stanley E. *Linguistic Analysis of the Greek New Testament: Studies in Tools, Methods and Practice*. Grand Rapids, MI: Baker Academic, 2015.

Porter, Stanley E., and Matthew Brook O'Donnell. "Theoretical Issues for Corpus Linguistics and the Study of Ancient Languages" in *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*, Łódź Studies in Language 8, edited by Andrew Wilson, Paul Rayson and Tony McEnery, 119-137. Frankfurt: Peter Lang, 2003.

Porter, Stanley E., and Andrew W. Pitts. *Fundamentals of New Testament Textual Criticism*. Grand Rapids, MI: Eerdmans, 2015.

Read-Heimerdinger, Jenny. *The Bezan Text of Acts: A Contribution of Discourse Analysis to Textual Criticism*. JSNTSup 236. Sheffield: Sheffield Academic Press, 2002.

Robinson, Maurice A. "Appendix: The Case for Byzantine Priority," in Maurice, A. Robinson, and William G. Pierpont. *The New Testament in the Original Greek: Byzantine Textform*. Southborough, MA: Chilton Book Publishing, 2005.

Robinson, Peter M. W. *Collate: Interactive Collation of Large Textual Traditions,* Version 2, Computer Program distributed y the Oxford Centre for Humanities Computing. Oxford, 2004.

Snyder, Julia A. *Language and Identity in Ancient Narratives*. Tübingen: Mohr Seibeck, 2014.

Wachtel, Klaus. "Editing the Greek New Testament no the Threshold of the Twenty-first Century", *Literary and Linguistic Computing* 15 (2000) 43-50.

Wachtel, Klaus. "The Corrected New Testament Text of Codex Sinaiticus" in *Codex Sinaiticus*: *New Perspectives on the Ancient Biblical Manuscript* edited by Scot McKendrick, David Parker, Amy Myshrall and Cillian O'Hogan, 97-106. London: The British Library, 2015.

Zuntz, Günther. *The Text of the Epistles: A Disquisition upon the Coprus Paulinum*. London: The British Academy, 1953.