A Corpus of Biblical Names in the Greek New Testament to Study the Additions, Omissions, and Variations across Different Manuscripts

Christoph Werner[1,*], Zacharias Shoukry[2], Soham Al-Suadi[2] and Frank Krüger[1]

[1]Hochschule Wismar – University of Applied Sciences, Philipp-Müller-Straße 14, 23966 Wismar, Germany
[2]University of Rostock, Universitätsplatz 1, 18055 Rostock, Germany

Abstract

The analysis of textual variants of verses in the New Testament across different manuscripts has mainly been done by close reading with manual effort. With the increasing number of transcriptions of the different manuscripts, quantitative analyses (so-called distant reading) can be used to search for patterns of omission, addition, or other variations, to formulate novel hypotheses to be investigated by close reading. In this work, we present a corpus of biblical names including spelling variation and inflections and their mentions in the transcriptions of the New Testament. By integrating and semantically enriching the data collected from different sources, we established a corpus that can be used for the quantitative study of omission, addition, and variation of such biblical names. To illustrate the corpus, we implement some use cases and show that well-known cases can be quantitatively reproduced. The corpus and all code are published under open licenses to enable reproduction, update, and maintenance.

Keywords

New Testament, Biblical Names, Textual Variation Units

1. Introduction

Research on the editions of the New Testament involves the study of textual variations across different manuscripts from several centuries and thus reflects the cultural background of such changes. Besides differences due to *small* grammatical variations, verses differ in their mention of biblical characters. For instance, additions, omissions, and other variations of biblical names can be observed, which are results of copy errors or selection due to cultural, gender, or other biases. The well known case of Junia(s) and Julia, for instance, where both names are used in different variations of the same verse, is subject of discussion in the field of textual criticism. The omission of Damarias in Acts 17:34 of the Codex Bezae is another case, leading to discussions about general gender biases of the manuscript itself.

With textual criticism, the above conflicting instances have been identified by *close reading*, the manual inspection and interpretation of the variations of verses across different manuscripts. Due to the long-lasting transcription efforts, for instance, by the Institute of New Testamental Textual Research (INTF) or the International Greek New Testament Project (IGNTP), the base

SemDH 2024: First International Workshop of Semantic Digital Humanities, May 26 or May 27, 2024, Hersonissos, Greece
christoph.werner@hs-wismar.de (C. Werner)
0009-0008-9907-251X (C. Werner); 0000-0002-9784-7034 (Z. Shoukry); 0000-0003-1098-208X (S. Al-Suadi); 0000-0002-7925-3363 (F. Krüger)


for automatic analyses have been established. In this work, we built upon the transcription efforts by integrating the textual data from both sources and further semantic enrichment. In particular, the contributions of this paper are: 1. By integration of different sources, we compiled a corpus of transcribed verses of the New Testament, which enables automated investigation of textual variations, 2. we compiled a dictionary of biblical names including their variations, by including grammatical inflection and other typical variations, 3. Finally, by analyzing omissions of biblical names in verses across different manuscripts, we illustrate the relevance of the data and quantitatively reproduce well known findings. In the following, we first give a short introduction to the New Testament, its history and, the source of verse variation. We then outline the data collection and processing and describe relevant characteristics of the generated corpus. Finally, we illustrate how the corpus can be used to generate hypotheses for further analyses based on closed reading.

2. History of Editions of the New Testament

If you open a Bible today, it is usually divided into two main parts, which have different headings depending on the edition. Common headings are "Old Testament" and "New Testament". This second part is a collection of 27 smaller writings that were most likely all written in Greek in the 1st–2nd century CE. In the 4th century, Jerome and others began to collect various Latin translations and compile a uniform, revised text from them, which resulted in the Vulgate, which became increasingly standardized and established over the course of the Middle Ages and was finally declared the authentic text by the Catholic Church in the 16th century. Erasmus of Rotterdam, who was guided by the humanist ideal "Ad fontes–To the sources", was also active during this period. He was not satisfied with a Latin translation, but wanted to return to the older Greek tradition. The problem, then as now, is that the autographs, i.e. the original papyri on which the New Testament texts were written, no longer exist. If we take all the Greek manuscripts known today from the 2nd–19th centuries together, we have around 5700,[1] and this figure does not include the thousands of manuscripts of translations into Latin, Coptic, Syriac, etc., not to mention the manuscripts of early church authors with biblical quotations, which are also considered textual witnesses. So we have all kinds of versions of the same texts, which naturally lead to variants that differ from one another. Almost all text-critical editions from the 16th–20th centuries were compiled by comparing manuscripts individually and noting the deviations by hand.

The New Testament (NT) manuscripts are classified into four distinct categories: papyri, majuscules, minuscules, and lectionaries. Papyri, being the oldest witnesses of the NT, often exist in fragmented form (see Figure 1). Majuscules, also referred to as biblical uncials, are characterized by their use of majuscule letters, which feature minimal ascenders and descenders. In contrast, minuscules are written in a small, cursive Greek script. Lectionaries can be encountered in both majuscule and minuscule Greek lettering styles. The Institute for New Testament Textual Research (INTF) at the University of Münster has cataloged all currently known manuscripts to the best of their ability in their New Testament Virtual Manuscript Room (NTVMR).

[1]This is an estimate from September 2023 by [1]

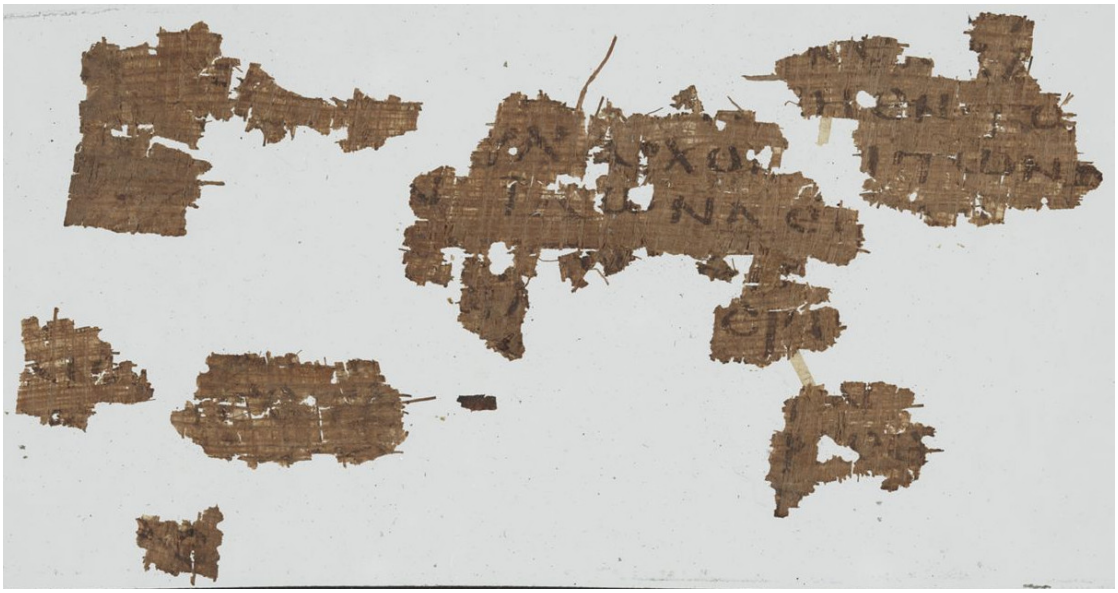


Figure 1: Fragmentation of papyrus P21 [2]

Various numbering schemes are employed for the above mentioned manuscripts.

With the Gregory-Aland Scheme (the de facto standard for biblical manuscript referencing), IDs differ depending on the manuscript type. Papyrus manuscripts are denoted by a Gothic/Black-letter P followed by a superscript number (e.g., $\mathfrak{P}^{52}$), often simplified to P52 for ease of display. Majuscules are identified by a leading zero followed by an incremental number (e.g., 0166). Minuscules are designated by an incremental document number alone. Lectionaries are indicated by a leading ℓ followed by an incremental number (e.g., ℓ2005). A capital L is often used in place of ℓ (e.g., L2005) due to display limitations.

The INTF Scheme uses a different approach. Instead of preceding letters ($\mathfrak{P}$, ℓ, or 0), it combines the document number with a leading digit indicating the manuscript type (1 for papyrus, 2 for majuscule, 3 for minuscule, and 4 for lectionary). Padding zeros are inserted between the leading digit and the document number – the document number is identical to its corresponding document number in the Gregory-Aland Scheme – to form a five-digit Document Identifier Number (docID). For instance, the Gregory-Aland notation $\mathfrak{P}^{52}$ is equivalent to 10052, indicating a papyrus manuscript. Similarly, the number 0166 is transformed into 20166, representing a Majuscule manuscript. In the case of the Gregory-Aland noted manuscript 365, it corresponds to 30365 in the INTF notation, denoting a minuscule manuscript. Lastly, ℓ2005 is the same as 42005, indicating a lectionary manuscript.

In this paper, the Gregory-Aland scheme is applied when mentioning or referencing manuscripts.

3. Data Collection and Processing

3.1. NTVMR Data

The New Testament Virtual Manuscript Room (NTVMR), managed by the INTF, offers an API[2] from which we retrieved docIDs of interest.

Cataloguing of the manuscripts has presumably been completed, but the numbering and its correction [1] is still the subject of discussion. The currently catalogued number of manuscripts, the docID ranges in use, next to the number of duplicates and merges forming the total known number of manuscripts are shown by manuscript category in Table 1. As one can see in Table 2 imaging, indexing, and transcribing is still a task in progress. Most progress in percentage terms has so far been made with the papyri, followed by the majuscules, minuscules, and lectionaries.

Manuscript Type	Catalogued	Ranges of docIDs in use	Removed/ Combined	Total
Papyri	142	10001–10142	6	136
Majuscules	332	20001–20326, 29994–29999	42	290
Minuscules	3,060	30001–33020, 39960–39999	159	2,901
Lectionaries	2,633	40001–42556, 49920–49975, 49979–49999	135	2,498

Table 1
Number of manuscripts catalogued by the NTVMR and their ranges of docIDs (as of 2024-03-05)

The API provides endpoints that can be used to access transcriptions[3] and metadata[4] for a specific manuscript respectively. Both request types require a docID to obtain a certain TEI XML file with transcription data or JSON file with metadata off the NTVMR server.

3.2. IGNTP Data

The International Greek New Testament Project (IGNTP) provides full transcriptions across selected manuscripts for John, Galatians, and Ephesians, whereas the transcriptions of Phillippians and 1 Corinthians are marked as 'in progress'. This data is accessible via direct downloads [3, 4] of zipped XML files. In addition to the transcriptions, the TEI files provided by the IGNTP also contain data on the respective manuscripts.

3.3. List of Names

As no machine read- and processable list of names in the Greek New Testament exists, it has been created in a largely manual and iterative process described in the following.

An initial compilation of biblical names from the New Testament was gathered from FactGrid. We sought data on all individuals mentioned in any of the books of the NT, resulting in a list of 305 biblical characters. It must be emphasized that biblical characters may share identical names, such as Mary of Bethany and Mary Magdalene. Given the theological debate surrounding such

[2] https://ntvmr.uni-muenster.de/community/vmr/api/metadata/liste/get/
[3] https://ntvmr.uni-muenster.de/community/vmr/api/transcript/get/?docID=<ID>&pageID=ALL&format=teiraw
[4] https://ntvmr.uni-muenster.de/community/vmr/api/metadata/manuscript/get/?docID=<ID>&format=json

Manuscript Type	Catalogued		Imaged		Indexed		Transcribed	
	count	%	count	%	count	%	count	%
Papyri	1,351	100	1,318	97.56	1,280	94.74	1,290	95.48
Majuscules	26,812	100	25,836	96.36	22,921	85.49	6,902	25.74
Minuscules	1,318,117	100	1,229,597	93.28	349,604	26.52	44,808	3.40
Lectionaries	802,998	100	412,644	51.39	20,570	2.56	3,315	0.41
Total	2,149,278		1,669,395		394,375		56,315	

Table 2
Statistics on processed manuscript pages by the NTVMR (data from API request to https://ntvmr.uni-muenster.de/community/vmr/api/statistics/pages/ as of 2024-03-05)

Manuscript Type	Number of Manuscripts	Removed/ Combined	Total	Number of Verses (by publisher)		
				IGNTP	INTF	ITSEE
Papyri	48	0	48	39	0	1,889
Majuscules	105	0	105	8,597	3002	24,324
Minuscules	348	2	346	103,236	0	82,948
Lectionaries	43	0	43	8,996	0	27,990

Table 3
Number of Manuscripts and Verses by Manuscript Type in the IGNTP Corpus (as of 2024-03-05)

potential identity overlap of certain biblical individuals, we opt to consolidate characters sharing the same name and subsequently refer exclusively to biblical names. The list of biblical names was compared to and expanded with information from [5], resulting in a total of 319 biblical names. Variations in grammatical cases were sourced from the Louw-Nida lexicon [6]. The subsequent search (described in section 3.6) led to an iterative refinement of alternative spellings, as we checked the verses marked as 'missing a name' for spelling variants of given name.

To facilitate later searches for names and their variations, it is imperative to compile a comprehensive list of all known spelling variations associated with each individual. This involves consolidating spelling variation of all grammatical cases into a list, as well as removing diacritics and transforming list entries into lowercase characters. The resulting list of lists encapsulates the diverse variants of individuals' names.

3.4. Parsing TEI Files for Transcription Data

Prior to parsing the previously acquired TEI files, a validity check is conducted. Through this process, a total of 41 of 1617 files (2.5%) are identified as invalid XML and do get excluded from parsing, and consequently from subsequent analysis. This is done to ensure a reasonable automation of the parsing process. Various factors contribute to the invalidity of XML files, such as discrepancies between opening and closing tags (n=7, 17%), undefined entities (n=10, 24%), duplication of attributes (n=1, 2.4%), junk after document element (n=1, 2.4%), as well as syntax errors resulting from invalid attribute names and/or values (n=22, 54%). It is pertinent to note that only files originating from the INTF exhibit non-valid status.

In essence, a TEI-XML file comprises a TEI header containing metadata about the document

and a text block containing transcription data. TEI markup [7] is utilized to represent the structural and semantic elements of the text, such as paragraphs, headings, lists, and quotations, using XML tags. For instance, tags delineate divisions like books or chapters, <ab> tags represent verses, and <w> tags denote words. Some words or entire parts may be unclear or missing, which are marked using <unclear> tags for ambiguous portions and <gap> tags for missing parts. Additionally, <supplied> tags indicate where known content has been inserted instead of setting a <gap> tag.

The transcription data follows a hierarchical structure based on the folio[5], organized by book, chapter, verse, and word. Since some verses span multiple pages, this hierarchy may be repeated several times within a single transcription document. Additionally, the same verse can appear multiple times on different folios of a document, particularly in lectionaries.

When processing the data verse by verse, we first link all related <ab> tags, where the values of the 'part' attribute ('I' for initial and 'F' for final) are decisive. If these attribute values occur in consecutive <ab> tags, these are to be combined. Otherwise, <ab> tags with the same 'name' attribute value are treated as individual verse transcriptions.

The verse blocks generated in this way are then searched for <w> tags and a list of the <w> tags found is created. This list is then parsed for text, resulting in a string representing the transcription. This string is then stripped of diacritics and formatted as lowercase letters.

As it later could be of importance which parts of the transcription have been marked as supplied or unclear, we generate a string of the same structure as the transcription string. In this string all characters have an initial value of 'c' (clear). By checking against the previously produced list of the <w> tags we are able to set unclear character values to 'u' and supplied characters to 's'. For example: the text string extracted from Listing 1 is seen in (1), the string which indicates the readability of the letters during transcription is seen in (2).

αμμιναδαβ δε εγεννησεν (1)

uuusuussc cc ccccccccu (2)

```
...
<ns0:w>
    <ns0:unclear>αμ</ns0:unclear>
    <ns0:lb break="no" />
    <ns0:unclear>μ</ns0:unclear>
    <ns0:supplied>ι</ns0:supplied>
    <ns0:unclear>να</ns0:unclear>
    <ns0:supplied source="na28" reason="illegible">δα</ns0:supplied>β</ns0:w>
<ns0:w>δε</ns0:w>
<ns0:w>εγεννησε<ns0:unclear>ν</ns0:unclear></ns0:w>
...
```

Listing 1: Example for clear, unclear, and supplied characters in the transcription of P1 by INTF

<gap> tags are not taken into account during parsing, as we are solely interested in the transcribed text. But for the sake of completeness those gaps should find their way into a later version of the data.

[5]In this context a folio is a manuscript page

Column	Description	Format	Example data
ga	Document Identifier by Gregory Aland Scheme	String	P1
bkv	Verse Identifier by BKV Scheme	String	B01K1V1
nkv	Verse Identifier by NKV Scheme	String	Matt.1.1
text	Transcription text	String	βιβλος γενεσεως ιυ χυ υυ δαυιδ υιου αβρααμ
marks	Marking of clear, unclear and supplied characters	String	cccccc cccccccc cc cc cc ccccc ssss cccccc
publisher	Transcription publisher	String	The Institut für neutestamentliche Textforschung
source	Download source of transcription	String	ntvmr

Table 4
Data Dictionary for Verses Collection

Extracted text and readability marks are saved alongside their docID and GA number, source, and publisher and verse identifier. It is to mention that 'source' describes the download source (either 'igntp' or 'ntvmr').

IGNTP and INTF use different forms of verse identifiers. The IGNTP bases its nomenclature on [8], but all separators are replaced by dots and spaces are removed (e.g., '1 Cor 1:3' becomes '1Cor.1.3'). The INTF instead assigns an ascending alphanumeric identifier to each book starting with B01 for the Gospel of Matthew and ending with B27 for the Book of Revelation. Whereby the numbering follows the listing of Books of the New Testament in [8] and similarly, chapters within a book are numbered with K and their verses with V (e.g., '1 Cor 1:3' becomes 'B07K1V3'). These verse identifiers are referenced below as BKV (INTF scheme) and NKV (IGNTP scheme). Since one of the two is always present, we are able to derive the other and save it alongside. All data extracted and generated during TEI parsing is saved to a data frame with the format depicted in Table 4.

3.5. Manuscript Metadata

During the parsing process of TEI files sourced from IGNTP and NTVMR, we successfully extracted essential metadata such as the GA number, docID, and occasionally a manuscript label. Further augmentation of this dataset was achieved through the incorporation of JSON files housing comprehensive metadata for each manuscript within the NTVMR. This supplementary information encompasses details such as docID, GA number (utilized for verse linkage), specifics on the location of storage (shelf instances), estimated period of origin, dimensions (both width and height), as well as counts of leaves, pages, columns, and lines. Notably, each page of a manuscript is accompanied by pertinent data regarding indexed content (verses on a page), hyperlinks to transcriptions and images, and indications of image protection necessitating NTVMR expert account authentication for viewing.

To further enrich our dataset with publicly accessible information, we conduct a query on dbpedia to acquire additional manuscript data. The query yields results comprising URIs,

Column	Description	Format	Example data
docID	Document Identifier by INTF scheme	Integer	30461
ga	Document Identifier by Gregory Aland Scheme	String	461
century	Century in Roman letters (with some exceptions being numeric)	String	835
pagesCount	Number of pages	String	688
leavesCount	Number of leaves	String	344
dbpedia	Link to dbpedia	String	http://dbpedia.org/resource/Uspenski_Gospels
label	manuscript name	String	Uspenski Gospels
source	Sources of data	String	ntvmr

Table 5
Data Dictionary for Manuscript Metadata Collection

manuscript labels, manuscript types and numbers, and temporal and spatial origins and/or discoverers of the manuscripts.

Upon scrutinizing the retrieved data, it became evident that manuscript numbers exhibit variation, being represented in distinct formats such as "𝔓48"@en, "''''𝔓24"@en, "ℓ2137"@en, and "ℓ 2144"@en or occasionally are presented solely as numeric values which do not give any clue on the type of manuscript. However, based on the RDF property 'form', the manuscript type is given as papyrus, uncial, minuscule, or lectionary.

To ensure uniformity and facilitate seamless data integration based on the respective GA number, a cleanup process was necessary. This involved removing all non-numeric characters from the manuscript number string. Subsequently, the remaining numerical value was concatenated with an initial character determined by the RDF property 'form', adhering to the GA notation convention.

After merging the different manuscript data sources we get a csv file with the columns described in Table 5

3.6. Search for Names

The main task involves the identification of occurrences and subsequent detection of omissions within the verses dataset built during TEI parsing.

Therefore we take the previously processed names, add a unique numeric nameID per name (for ease of later use), explode the list of variations and add a unique numeric variantID to each generated entry of variants (see Table 6).

After retrieving the unique BKV verse identifiers from the list of verse transcriptions (see Table 4 for its data dict), the process of searching is parallelized in that manner that all name variants are searched on a BKV-by-BKV basis.

This parallelized approach involves filtering a copy of all verse transcriptions for entries corresponding to a given BKV verse identifier. Subsequently, each transcription text field is scanned for all name variants. Upon identification of a variant, the associated nameID is

Column	Description	Format	Example data
label:en	Name in English	String	Aaron, brother of Mose
gender	Genus of the person	String	m
label:el	Name in Greek	String	ααρων
factgrid	FactGrid ItemID	String	Q165847
variant	Spelling variant	String	ααρωνος
wordID	Unique word identifier	Integer	4
variantID	Unique variant identifier	Integer	7

Table 6
Data Dictionary for Name Collection

Column	Description	Format	Example data
ga	Document Identifier by GA scheme	Integer	10001
bkv	Verse Identifier by BKV Scheme	String	B01K1V1
text	Transcription text	String	βιβλος γενεσεως ιυ χυ υυ δαυιδ υιου αβρααμ
wordID	Unique word identifier	Integer	23
variantID	Unique variant identifier	Integer	77
occurrence	Indicator of occurrence	Boolean	True

Table 7
Data Dictionary for Name Occurrences Collection

appended to a set specific to that verse transcription (denoted as 'found'), representing all detected names within. Additionally, a BKV-specific set (denoted as 'occurrences') is updated to include all names detected in any text associated with the given BKV. By comparing the 'found' set against the 'occurrences' set for the respective BKV, we can ascertain the names missing from each transcription.

Afterwards, the found and missing sets do get exploded separately for each verse transcription. With this, the data frame now contains rows with information on a certain verse and the occurrence or omission of one specific name in it. The corresponding data dict for the described data frame is given in Table 7.

4. Analysis of the Transcription Corpus

4.1. Transcription Overview

At first glance at the transcription corpus, we can see that the INTF has the most transcribed verses on papyri, minuscules, and majuscules as well as in total. Followed by the Institute for Textual Scholarship and Electronic Editing (ITSEE) and IGNTP.

As a result of the multi-source character of our transcription corpus, which incorporates transcriptions from IGNTP and INTF, there are duplicates of verse transcriptions to be expected. When considering duplicates as entries with identical docID/bkv combinations, we identify 3817 instances. However, to assess whether these duplicates are also identical in the transcribed text, we examine entries with identical docID/bkv/text combinations, revealing 3624 duplicates.

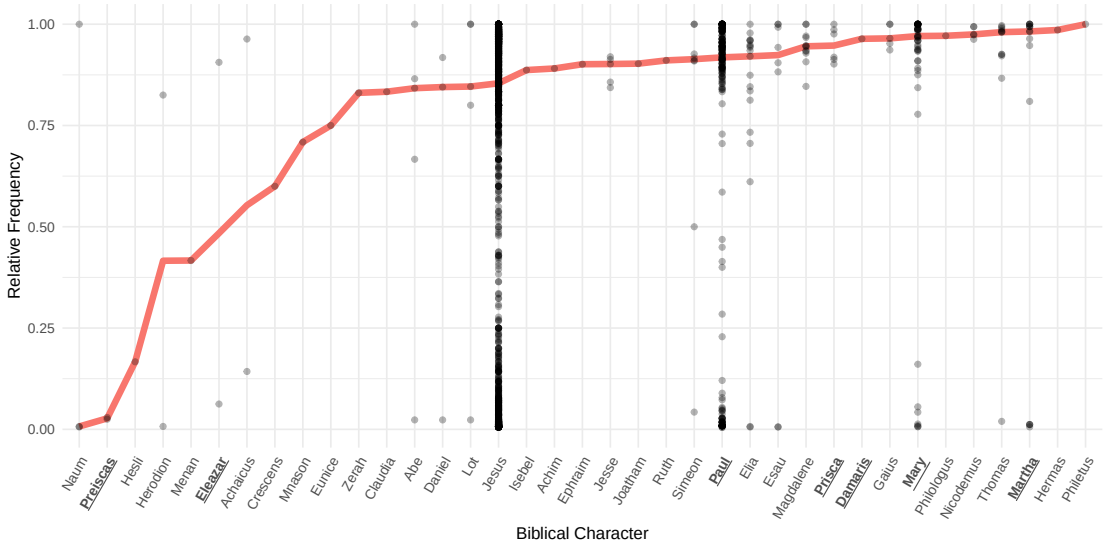


Figure 2: Relative occurrence frequency of a selected subset of biblical characters across all verses. The red line depicts the median (excluding 0s). Characters are ordered by the median relative occurrence frequency.

Duplicate transcriptions remain in the corpus for the sake of completeness of the collection of transcriptions.

4.2. Analysis of Additions, Omissions, and Variations of Names

To illustrate the value of the corpus presented in this work, in the following, we illustrate some use cases. To this end, we first analyze omissions of names, by computing the relative frequency of the occurrence of a biblical character across all variations of a verse. Figure 2 depicts a subset of biblical characters including their relative occurrence frequency within different verses. For each verse and character, the frequency was determined by the ratio of manuscripts where the character was included in a verse and the overall number of manuscripts that actually contain this verse. Frequencies of 0.0 were left out, as they represent verses where the particular character was never included. From the figure, several observations can be made, some of which are summarized in the following.

Firstly, for Esau one verse with a relative occurrence frequency of 1 is evident, indicating that it is included in all variations of this verse. Secondly, Eleazar has one high ($106/117 = 91\%$) and one low ($1/16 = 6\%$) relative occurrence frequency reflecting different omission resp. variation pattern. While different other patterns can be observed from the figure that suggest omissions, additions, or variations, a closer look at particular verses is necessary to draw reliable conclusions. Figure 3 illustrates the occurrence of different biblical names across different variations of a particular verse. In the following, these patterns are analyzed more closely.

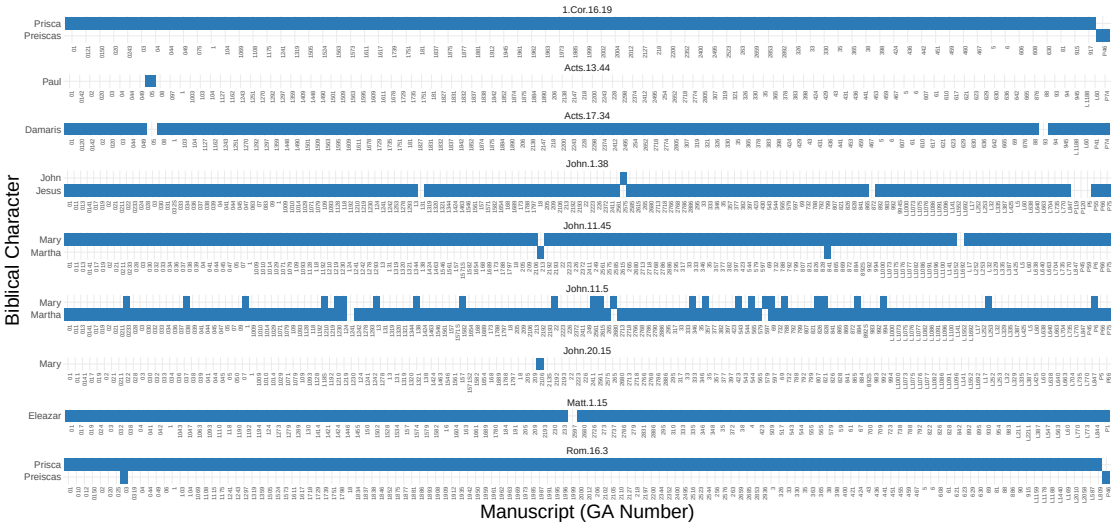


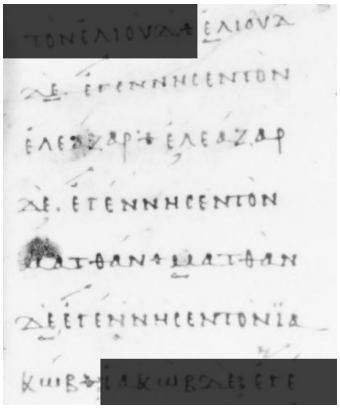
Figure 3: Occurrences of different biblical names across different variations of the same verse. A blue box indicates the inclusion of the name in the verse within a particular manuscript.

4.3. Examples of Additions, Omissions, and Variations of Names

During the iterative search and the follow-up analysis of name variations, we found some known and possible unknown examples of omissions, additions, and variations which we show below.

We found Paul to be inserted in Acts 13:44 by Codex Bezae (majuscule 05). This finding is consistent with both, the current text-critical hand editions NA28 [9] and ECM [10]. On the other hand, the woman Damaris is omitted from Codex Bezae in Acts 17:34, which is also evident from NA28 and ECM. In minuscule 1 we have found a potentially feminine variant of the name Epaphroditus in Phil 2:25 namely επαφροδιτα. The mentioned name επαφροδιτα is probably a copyist's error, since the immediate context does not allow for a feminine interpretation (the apposition is τον αδελφον, which is a grammatically clearly masculine word, namely "the brother"). We were also able to locate the already known and in [11] discussed variant of the forms Prisca and Preiscas in Rom 16:3 (P46 and 03) and 1 Cor 16:19 (P46). Another finding is the simultaneous occurrence of Prisca and Preiscas in 03, as corrections were made in a manuscript. We can confirm that Martha (e.g., P66) and Mary (e.g., 038 resp. Θ, Codex Koridethi) are interchanged in John 11:5, which has been examined before [12] and after [13] the publication of NA28. We can also confirm the variation of Mary and Martha (attested in 213, noted in the current preliminary online ECM of John) in John 11:45. Yet another variant is found in minuscule 841, which speaks of both Mary and Martha. We also found a striking variation in the minuscule 2575 in John 1:38 which has, to our knowledge, not yet been taken into account in textual criticism: John is written here instead of Jesus. Further, we were able to locate yet another omission in Matt 1:15, where Eleazar (ελεαζαρ) is skipped in the minuscule 2597. This verse is part of a genealogy in which copyists have probably slipped in the line with their eyes, because exactly the same verb form and article (homoioteleuton) appears between

father and child each time (“Eliud begat Eleazar. Eleazar begat Mattan. Mattan begat Jacob”) like it can be seen in Figure 4. A curious case is the addition of Mary in Jesus’s direct speech in John 20:15 (minuscule 2106). Here Jesus addresses Mary by her name, which is not evident in any other manuscript off our dataset. On further inspection, this addition might have happened as another case of one copyist’s slipping in the line of text, as the direct speech is introduced in both verses with exactly the same words (“Jesus said to her”).



(a) Handwriting of Matt 1:15 in Majuscule 045 [14]

(b) Transcription of Matt 1:15 in Majuscule 045
ελιουδ
δε εγεννησε τον
ελεαζαρ ελεαζαρ
δε εγεννησε τον
ματθαν ματθαν
δε εγεννησε τον ια
κωβ

(c) Transcription of Matt 1:15 in minuscule 2575
ελιουδ
δε εγεννησε τον
⋮
⋮
ματθαν ματθαν
δε εγεννησε τον ια
κωβ

Figure 4: Depiction of how a line slip due to a homoioteleuton could have happened: line 2 and line 4 are identical

5. Related Work

A work similar to ours is [15], which presents a registry of Hebrew names and an analysis of name occurrences within the lists found in the Torah book of Ezra–Nehemiah. Subsequently, [16] utilized this registry to establish the “Ancient Hebrew Personal Names” database.

We have found a deficiency in the accessibility of a thorough, machine-readable, or queryable compilation of names found within the Greek New Testament. Although efforts, such as those by FactGrid, have been made to compile such lists, they often lack completeness, Greek spelling of names, variations in name spelling, or comprehensive coverage across manuscripts. Our dataset is positioned to complement existing initiatives with entries addressing these shortcomings.

Furthermore, discussions have emerged regarding the downplay of females [17], debates concerning the gender attribution of certain names [18] [19], and inquiries into the textual traditions containing additions such as Martha of Bethany [13]. These discussions highlight the complexities surrounding omissions, additions, and variations in name usage, warranting further scholarly attention, in which our data can be of use.

When it comes to the analysis of textual variation in revision histories, some work has been done in the context of Wikipedia, where the main focus was the identification of vandalism and biased statements based on information about the corresponding editor. For instance, [20] identifies different revision patterns on a set of almost 7000 Wikipedia article revisions.

6. Limitations

The transcription process is not yet automated and will probably remain largely manual work in the future. This makes it all the more important that transcribers adhere to certain rules and guidelines like [21] and [22] to maintain conformity and reusability, as well as guarantee completeness and accuracy. However, precisely this reusability is currently a problem, as different transcribers do not fully adhere to the above guidelines, and the guideline version used is not mentioned in the transcript files. As a result, for example, it is often not indicated from which source a 'supplied' letter originates, or why a gap occurs in the text.

As we conduct a string search for names on the transcription corpus, we get a certain amount of false positives, which result in falsely negative entries in the list of occurrences. For example, there is the accusative form of Zeus ('διᾰ') which generates a lot of such entries, as it is also a preposition (in English: via, by, for, into, over, to). One could now argue that certain verses which show this should be excluded from the name search. However, we want to include all possible changes, including additions/omissions/variations that may only occur in one verse. For this reason, we keep these entries in the dataset for later analysis and removal.

While the string search does allow the disambiguation of different (or the same) persons in general, the corpus described in this paper, enables establishing hypotheses about the usage of different names for the same person and the subsequent quantitative analysis of mention patterns, for instance, by correlation. This, for example, is the case on Prisca and Preiscas, as there are discussions on a certain spelling being both feminine and masculine and whether the supposedly masculine form of the name could be just another spelling variation of the feminine form.

7. Conclusions and Future Work

In this paper, we present a novel data set which was collected by manually integrating and semantically enriching different public data sources for the study of omissions, additions, and variations of biblical names in the Greek New Testament. To this end, we illustrate the diverse origins of transcriptions pertaining to Ancient Greek New Testament manuscripts and describe the process of compiling transcriptions in detail. With the presented corpus of transcriptions and occurrences of names, a step has been taken towards the automatic creation of hypotheses for textual criticism. For instance, by correlating patterns of occurrences of names of apparently different characters, the established corpus allows to investigate name variations exceeding plain grammatical variants. We were able to show already known additions, omissions, and variations of name occurrences in our data based on well known examples from the literature. Moreover, we discovered a not yet discussed case of an addition of Mary in John 20:15.

While in its current form, the corpus can be used for the illustrated analyses, we are aware of some limitations, including false positives resulting from similarities between names and prepositions. To address this issue we plan to utilize methods of machine learning such as: Part Of Speech (POS) Tagging and Named Entity Recognition (NER). This requires a manual annotation for training and evaluation, but also needs particular attention to the drawbacks of such methods. To make the data semantically meaningful, accessible, and explorable for

further research, we plan to create a knowledge graph from the provided dataset for the FAIR publication. However, to the best of our knowledge, currently there are no ontologies for the semantic description of biblical names and characters in the New Testament. To this end, we plan to develop an appropriate ontology.

Used Software, Data, and Code Repositories

We have used Python (v3.12.1) and R (v4.3.1) for retrieving, processing, analyzing, and plotting data. Notable packages in use are: beautifulsoup4 (v4.12.3), jupyter (v1.0.0), notebook (v7.0.6), pandas (v2.1.4), sparqlwrapper (v2.0.0), as well as tidyverse (v2.0.0).

Data generated during this project is made available [23] on Zenodo. The Sourcecode of this project is published on GitHub[6].

Acknowledgments

We would like to thank Jan Krans-Plaisier and Peter-Ben Smit for their help in familiarising us with the topic and for their insights. Additionally, we thank Corinna Stratmann for her help in browsing dictionaries in search of relevant entries. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) 513300936.

References

[1] K. Leggett, G. S. Paulson, How Many Greek New Testament Manuscripts Are There REALLY? The Latest Numbers, 2023. URL: https://ntvmr.uni-muenster.de/intfblog/-/blogs/how-many-greek-new-testament-manuscripts-are-there-really-the-latest-numbers.

[2] Special Collections and Archives, Trexler Library. Muhlenberg College, P. Oxy. 1227: St. Matthew's gospel, xii., online, 2015. URL: https://library.artstor.org/#/asset/SS7730556_7730556_9313349, accessed 2024-03-11.

[3] The Principio Project, The International Greek New Testament Project, Papyri, Majuscules, Minuscules, and Lectionaries of John, online, 2024. URL: https://itseeweb.cal.bham.ac.uk/iohannes/transcriptions/, accessed 2024-03-04.

[4] Institute for Textual Scholarship and Electronic Editing Birmingham, Electronic Resources for the Textual Tradition of the Epistles of Paul, online, 2023/2024. URL: https://itseeweb.cal.bham.ac.uk/epistulae/, accessed 2024-03-04.

[5] W. Bauer, Griechisch-deutsches Wörterbuch zu den Schriften des Neuen Testaments und der frühchristlichen Literatur, 6., völlig neu bearbeitet auflage ed., Walter de Gruyter, Berlin, 2012. Frühere Auflage unter dem Titel: Bauer, Walter: Griechisch-deutsches Wörterbuch zu den Schriften des Neuen Testaments und der übrigen urchristlichen Literatur.

[6] J. P. Louw, E. A. Nida, Greek-English Lexicon of the New Testament, volume 1, United Bible Societies, New York, 1988.

[6] https://github.com/chr-werner/SemDH2024-GreekNewTestamentNames

[7] Text Encoding Initiative, TEI: Guidelines for Electronic Text Encoding and Interchange, P5 Version 4.7.0., revision e5dd73ed0, online, 2023. URL: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html, accessed 2024-03-04.

[8] J. F. K. Billie Jean Collins, Bob Buller, The SBL Handbook of Style, SBL Press, 2014. doi:10.2307/j.ctt14bs6ct.

[9] B. Aland, K. Aland, J. Karavidopoulos, C. M. Martini, B. M. Metzger (Eds.), Novum Testamentum Graece, 28 ed., Deutsche Bibelgesellschaft, Stuttgart, 2012.

[10] H. Strutwolf, G. Gäbel, A. Hüffmeier, G. Mink, K. Wachtel (Eds.), Die Apostelgeschichte: The Acts of the Apostles: Novum Testamentum Graecum: Editio Critica Maior, Deutsche Bibelgesellschaft, Stuttgart, 2017.

[11] D. A. Kurek-Chomycz, Is There an "Anti-Priscan" Tendency in the Manuscripts? Some Textual Problems with Prisca and Aquila, Journal of Biblical Literature 125 (2006) 107. doi:10.2307/27638349.

[12] K. von Tischendorf, Novum Testamentum Graece: Ad Antiquissimos Testes Denuo Recensuit Apparatum Criticum Omni Studio Perfectum Apposuit Commentationem Isagogocam, volume I of *Editio Octava Critica Maior*, Giesecke & Devrient, Leipzig, 1869.

[13] E. Schrader, Was Martha of Bethany Added to the Fourth Gospel in the Second Century?, Harvard Theological Review 110 (2017) 360–392. doi:10.1017/s0017816016000213.

[14] Library of Congress, Collection of Manuscripts – Monastery of Dionysios 55. (old 10). (Greg. 045, Ω). Four Gospels. 8th/9th cent. 259 f., 2024. URL: https://www.loc.gov/resource/amedmonastery.00271050008-ma/?sp=12&r=0.51,0.18,0.177,0.153,0, accessed 2024-03-04.

[15] A. Frank (Ed.), Asaf, Juda, Hatifa - Namen und Namensträger in Esra/Nehemia, number 78 in Stuttgarter Biblische Beiträge (SBB), Verlag Katholisches Bibelwerk, Stuttgart, 2020.

[16] A. Frank, H. Rechenmacher, Morphologie, Syntax und Semantik Althebräischer Personennamen, Universitätsbibliothek der Ludwig-Maximilians-Universität München, 2020. doi:10.5282/UBM/EPUB.73364.

[17] R. G. Fellows, Early Textual Variants That Downplay the Roles of Women in the Bethany Account, Textual Criticism 28 (2023) 67–82.

[18] E. J. Epp, Junia: The First Woman Apostle, Fortress Press, Minneapolis, 2005.

[19] R. G. Fellows, Early Sexist Textual Variants, and Claims That Prisca, Junia, and Julia Were Men, The Catholic Biblical Quarterly 84 (2022) 252–278.

[20] Z. Ma, J. Tao, J. Hu, The dynamics of wikipedia article revisions: an analysis of revision activities and patterns, International Journal of Data Mining, Modelling and Management 9 (2017) 298. doi:10.1504/ijdmmm.2017.088415.

[21] H. Houghton, C. Smith, IGNTP guidelines for XML transcriptions of New Testament manuscripts (version 1.6), 2023. URL: http://epapers.bham.ac.uk/4301/.

[22] A. C. Myshrall, R. Kevern, H. Houghton, IGNTP guidelines for the transcription of manuscripts using the Online Transcription Editor, 2020. URL: http://epapers.bham.ac.uk/3436/.

[23] C. Werner, F. Krüger, Z. Shoukry, S. Al-Suadi, A Corpus of Biblical Names in the Greek New Testament to Study the Additions, Omissions, and Variations across Different Manuscripts, 2024. doi:10.5281/zenodo.10985520.