ON THE IMPACT OF TIME PROXIMITY ON THE ALIGNMENT OF SPELLING VARIANTS IN HISTORICAL ENGLISH BIBLES: A CASE STUDY

Maria Moritz

Jan. 25-26 2018, *Corpus-based Research in the Humanities, Vienna, Austria*

eTRAP
Electronic Text Reuse Acquisition Project

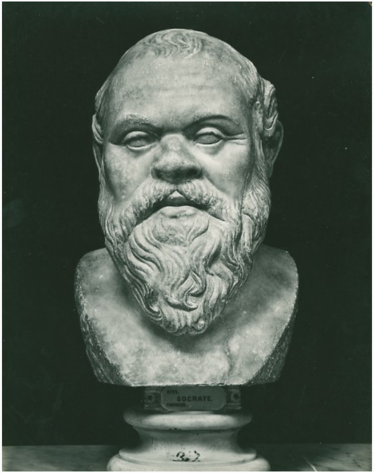
GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

TABLE OF CONTENTS

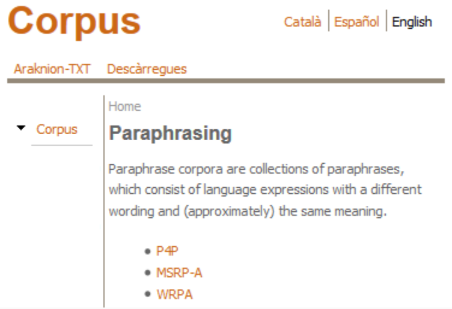
HISTORICAL TEXT REUSE DETECTION

HISTORICAL TEXT REUSE DETECTION

- Text Reuse: Written repetition of text, e.g., quotations, allusions, translations
- Useful in: Phylogenetics, Fragmentary Authors (Socrates->Plato)
- Modern use-case: Plagiarism detection



PARAPHRASTICALITY AS CHALLENGE



Paraphrasing

Paraphrase corpora are collections of paraphrases, which consist of language expressions with a different wording and (approximately) the same meaning.

- P4P
- MSRP-A
- WRPA

... with trade totaling more than $34 billion.

... with trade volume of $33.4 billion last year.

- Greek plane lands at UK airport after dire warning.
- A bomb threat has prompted a Greek Olympic Airlines passenger plane to make an emergency landing, escorted by British Tornado jets, at London's Stansted Airport.

[1] [2]

[1] http://pan.webis.de/

[2] http://clic.ub.edu/corpus/en/paraphrases-en

ANALYSIS OF MODIFICATIONS IN PARAPHRASES

In historical texts, we encounter even stronger challenges, due to:

- strong variation during long transmission time
- incomplete witnesses
- diverse reuse types

To reinforce research in the field, we want to:

- investigate how a text is modified
- to understand the broader context of the reuse happening

Our long-term goal is to build a formalism behind the transformation (modification) of reuse.

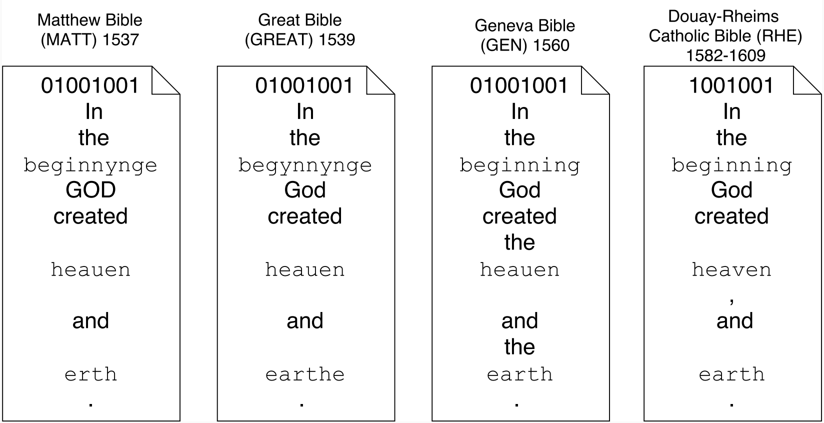
image: https://nieuws.kuleuven.be/en/content/2015/ku-leuven-restores-and-displays-ancient-manuscripts-from-timbuktu

STUDY DESIGN

STUDY DESIGN – OVERVIEW

We use a monolingual, diachronic corpus of English Bibles.

- We investigate if time proximity can help to map historical writing variants among each other using a simple character-distance measure.



Matthew Bible (MATT) 1537	Great Bible (GREAT) 1539	Geneva Bible (GEN) 1560	Douay-Rheims Catholic Bible (RHE) 1582-1609
01001001	01001001	01001001	1001001
In	In	In	In
the	the	the	the
beginnynge	begynnynge	beginning	beginning
GOD	God	God	God
created	created	created	created
		the	
heauen	heauen	heauen	heaven
			,
and	and	and	and
		the	
erth	earthe	earth	earth
.	.	.	.

STUDY DESIGN – RESEARCH QUESTIONS

We seek to find out:

1. RQ) Does the use of temporally-close Bibles improve the alignment of historical writing variants?
2. RQ) Whether and how does time proximity in historical texts help to normalize old variants of text to modern spelling?, and
3. RQ) What are specific problems to align a historical Bible corpus?

STUDY DESIGN - METHODOLOGY

We define operations to model modifications in text.

operation verbose	operation name
perfect match	**NOP**(word1,word2)
lower-casing matches	**lower**(word1,word2)
lemmatizing matches	**lem**(word1,word2)
short levenshtein matches	**lev**(word1,word2)
words are synonyms	**syn**(word1,word2)
word1 is hypernym of word2	**hyper**(word1,word2)
word1 is hyponym of word2	**hypo**(word1,word2)

STUDY DESIGN - DATA

- We collect English Bible translations:
    1. Parallel Text Project[3]
    2. Mysword[4]
    3. Bible Study Tools[5]
- Historical Bibles ranging from 1500s to 1900
- Excluding literal translations (e.g., Young's, Smith's), because of vocabulary diversity
- Exclude Darby Bible (1890) for the above reason, and because it is influenced by translations in other languages

[3] http://paralleltext.info

[4] http://mysword.info/

[5] https://www.biblestudytools.com/

STUDY DESIGN - DATA

- MATT, GREAT and GEN are is written in EME with words appearing & being spelled different than today(e.g., “daye”, “deuyde”, and “heaue”.
- MorphAdorner can normalize words such as “catell” (GREAT), “likenes” (MATT),
- but “lycknesse” (MATT), “licknesse” (GREAT) remain untouched.
- The remaining Bibles contain words that end in “-eth” (archaic), e.g., creepeth, yieldeth.

Bible	**date**
Matthew Bible (MATT)	1537
Great Bible (GREAT)	1539
Geneva Bible (GEN)	1560
Douay-Rheims Catholic Bible (RHE)	1582-1609
Douay-Rheims Challoner Revision (DRC)	1749-1752
King James (KJV)	1611-1769
The Webster Bible (WBT)	1833
English Revised Version (ERV)	1881-1894

DATA ALIGNMENT

DATA ALIGNMENT - PRE-PROCESSING

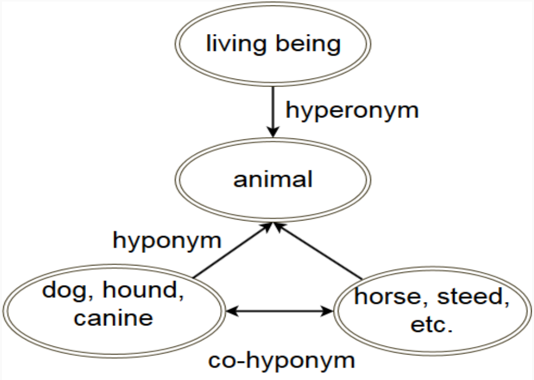
- We use MorphAdorner[6] to tokenize and lemmatize the text.
- MorphAdorner works list-, and rule-based, using Porter Stemmer

token	pos tag	normalized	lemma
1001003	crd	1001003	1001003
TAB	n1	TAB	tab
Than	cs	Than	than
God	np1	God	God
sayd	vvd	said	say
let	vvb	let	let
there	pc-acp	there	there
be	vbi	be	be
light	j	light	light
&	cc	&	and
there	a-acp	there	there
was	vbds	was	be
lyght	vvi	light	light
LINE	n1	LINE	line

[6] http://morphadorner.northwestern.edu

DATA ALIGNMENT – SYNSET DICTIONARY

We query the lemmas in BabelNet API to find synonym, hypernym, hyponym, and cohyponym relations between the words of two verses (Navigli et al. 2012)



DATA ALIGNMENT

We define operations to model modifications in text. We apply these operations in a prioritized order.

operation verbose	operation name
perfect match	**NOP**(word1,word2)
lower-casing matches	**lower**(word1,word2)
lemmatizing matches	**lem**(word1,word2)
short levenshtein matches	**lev**(word1,word2)
words are synonyms	**syn**(word1,word2)
word1 is hypernym of word2	**hyper**(word1,word2)
word1 is hyponym of word2	**hypo**(word1,word2)

DATA ALIGNMENT – RESULTS

		known lemmas (*lem*)			newly found edits (*lev*)		
source Bible	**target Bible**	**source types**	**target types**	**tokens**	**source types**	**target types**	**tokens**
MATT	GREAT	8,595	7,939	110,779			
GREAT	GEN	7,531	6,105	147,671			
GEN	RHE	5,300	4,534	115,027			
RHE	DRC	392	406	777			
DRC	KJV	2,713	2,747	24,206			
KJV	WBT	706	717	7,242			
WBT	ERV	1,734	1,816	11,908			

- Our distance measure “lev” fuzzily matches 2/7 characters with min length of 6.
- It works especially well for mapping proper names, e.g. Hyerusalem & Ierusalem.
- We align about half as many types with “lev” compared to the types that are aligned after lemmatization.
- Alignment between RHE-DRC and KJV-WBT is esp. unspectacular, because the target is revision of its predecessor.

DATA ALIGNMENT – RESULTS

		known lemmas (*lem*)			newly found edits (*lev*)		
source Bible	target Bible	source types	target types	tokens	source types	target types	tokens
MATT	GREAT	8,595	7,939	110,779	4,683	4,508	9,795
GREAT	GEN	7,531	6,105	147,671	3,178	2,753	9,359
GEN	RHE	5,300	4,534	115,027	1,471	1,424	6,296
RHE	DRC	392	406	777	349	359	1,212
DRC	KJV	2,713	2,747	24,206	1,235	1,199	4,316
KJV	WBT	706	717	7,242	594	592	2,233
WBT	ERV	1,734	1,816	11,908	974	958	2,772

- Our distance measure "lev" fuzzily matches 2/7 characters with min length of 6.
- It works especially well for mapping proper names, e.g. Hyerusalem & Ierusalem.
- We align about half as many types with "lev" compared to the types that are aligned after lemmatization.
- Alignment between RHE-DRC and KJV-WBT is esp. unspectacular, because the target is revision of its predecessor.

DATA ALIGNMENT – RESULTS

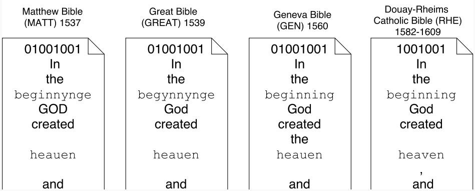
		known lemmas (*lem*)			newly found edits (*lev*)		
source Bible	target Bible	source types	target types	tokens	source types	target types	tokens
MATT	GREAT	8,595	7,939	110,779	4,683	4,508	9,795
GREAT	GEN	7,531	6,105	147,671	3,178	2,753	9,359
GEN	RHE	5,300	4,534	115,027	1,471	1,424	6,296
RHE	DRC	392	406	777	349	359	1,212
DRC	KJV	2,713	2,747	24,206	1,235	1,199	4,316
KJV	WBT	706	717	7,242	594	592	2,233
WBT	ERV	1,734	1,816	11,908	974	958	2,772
sum		16,311	15,094	417,610	10,587	9,915	35,983
MATT	ERV	8,137	5,317	181,451	2,682	2,160	8,561

- Our distance measure “lev” fuzzily matches 2/7 characters with min length of 6.
- It works especially well for mapping proper names, e.g. Hyerusalem & Ierusalem.
- We align about half as many types with “lev” compared to the types that are aligned after lemmatization.
- Alignment between RHE-DRC and KJV-WBT is esp. unspectacular, because the target is revision of its predecessor.

DATA ALIGNMENT – RESULTS

Variant Dictionary

- 5,803 entries containing types that result from the alignment
- Key: first appearance of a word that closes an alignment chain, i.e., word of "youngest" Bible
- Values: all other types of words that appear in one or more alignment chains according to a key



Matthew Bible (MATT) 1537	Great Bible (GREAT) 1539	Geneva Bible (GEN) 1560	Douay-Rheims Catholic Bible (RHE) 1582-1609
01001001	01001001	01001001	1001001
In	In	In	In
the	the	the	the
beginnynge	begynnynge	beginning	beginning
GOD	God	God	God
created	created	created	created
		the	
heauen	heauen	heauen	heaven
			,
and	and	and	and

- ***offering**: offreth offeryng offring offereth offeringe offer offered offred offerynge offrynges offryng offerings offrynge*
- ***vineyard**: venyarde vynearde vineyarde vyneyarde vineyards vyneyardes vyneyard vyneard vineiarde vyneiarde viniyardes vineyardes vineiard*

DATA ALIGNMENT – ERROR CLASSIFICATION

We differ:

- WordNet,
- Pre-Processing, and
- AUXialiary errors

Example from 19-057-003 Psalm 57:3

source	swalowe	my	Selah	for	faythfulnes	shall	wold
target	eate	me	Sela.	forth	treuth	will	would
error class	WN	recall	PP	recall	WN	AUX	recall

DATA ALIGNMENT – ERROR CLASSIFICATION

We manually evaluate ten randomly picked verses from each Bible alignment pair (70 verse, ca. 1400 tokens).

Bible		lem alignments		lev alignments			error types		
source	target	correct	wrong	true pos	false pos	false neg	WN	PP	AUX
MATT	GREAT	32	0	2	0	3	3	2	0
GREAT	GEN	56	1	0	0	4	1	2	2
GEN	RHE	33	0	1	0	0	0	0	2
RHE	DRC	2	0	0	0	0	0	0	0
DRC	KJV	5	0	0	0	0	1	0	2
KJV	WBT	1	0	0	0	0	0	0	0
WBT	ERV	7	0	1	0	0	0	0	0

CONCLUSION

CONCLUSION AND NEXT STEPS

Summary

- Alignment of historical variants is an prerequisite for analyzing modifications in text reuse.
- The extra *lev* operation improves its alignment by about 50% as many types as SOTA lemmatizers do.

Future Work

- Combine statistical alignment and operation-based alignment
- Expand the approach to collect variants among all "temporal" directions
- Use derivation dictionaries to align words with different POS
- Proper lemma matching needs further investigation

REFERENCES

- Archer, Dawn, McEnery, Tony, Rayson, Paul, and Hardie, Andrew (2003): Developing an automated semantic analysis system for early modern english. In: *Corpus Linguistics 2003 conference*.
- Baron, Alistair, and Rayson, Paul (2008): Vard2: A tool for dealing with spelling variation in historical corpora. In: *Postgraduate conference in corpus linguistics*.
- Burns, Philip R (2013): Morphadorner v2: A java library for the morphological adornment of English language texts. `http://morphadorner.northwestern.edu`. [Acc. Jan. 2018]
- DeNero, John, and Klein, Dan (2007): Tailoring word alignments to syntactic machine translation. In: *Proceedings of the Annual Meeting on Association for Computational Linguistics*, volume 45.
- Levenshtein, Vladimir I (1966): Binary codes capable of correcting deletions, insertions, and reversals. Doklady Akademii Nauk SSSR, 163(4), 1965. (1966) Russische, Englische Übersetzung. In: In: *Soviet Physics Doklady*, Vol. 10, No. 8.
- Marlowe, Michael (2017): John Nelson Darby's Version. `http://www.bible-researcher.com/darby.html`. [Acc. Nov. 2017]
- Mayer, Thomas, and Cysouw, Michael (2014): Creating a massively parallel bible corpus. In: *Proceedings of LREC'14*. European Language Resources Association (ELRA).
- Moritz, Maria, Wiederhold, Andreas, Pavlek, Barbara, Bizzoni, Yuri, and Büchler, Marco (2016): Non-literal text reuse in historical texts: An approach to identify reuse transformations and its application to bible reuse. In: *Empirical Methods in Natural Language Processing (EMNLP'16)*, Austin, TX, USA. ACL.
- Navigli, Roberto, and Ponzetto, Simone Paolo (2012): Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*
- Riversoft Systems. Mysword. `www.mysword.info/`, 2011–2017.
- Bible Study Tools (2017): Bible study tools. `http://www.biblestudytools.com/`. [Jan. 2018].
- Yang, Yi, and Eisenstein, Jacob (2016): Part-of-speech tagging for historical english. *CoRR*, abs/1603.03144.

Thank you!

DATA ALIGNMENT – RESULTS

		known lemmas (*lem*)			newly found edits (*lev*)		
source Bible	target Bible	source types	target types	tokens	source types	target types	tokens
MATT	GREAT	8,595	7,939	110,779	4,683	4,508	9,795
GREAT	GEN	7,531	6,105	147,671	3,178	2,753	9,359
GEN	RHE	5,300	4,534	115,027	1,471	1,424	6,296
RHE	DRC	392	406	777	349	359	1,212
DRC	KJV	2,713	2,747	24,206	1,235	1,199	4,316
KJV	WBT	706	717	7,242	594	592	2,233
WBT	ERV	1,734	1,816	11,908	974	958	2,772
sum		16,311	15,094	417,610	10,587	9,915	35,983
MATT	ERV	8,137	5,317	181,451	2,682	2,160	8,561

DATA ALIGNMENT – STATISTICAL ALIGNMENT

Bible		lem alignments		lev alignments			other operations			error types		
source	target	correct	wrong	true pos	false pos	false neg	syn	hyper	hypo	WN	PP	AUX
MATT	GREAT	32	0	2	0	3	2	1	0	3	2	0
GREAT	GEN	56	1	0	0	4	2	2	0	1	2	2
GEN	RHE	33	0	1	0	0	9	0	3	0	0	2
RHE	DRC	2	0	0	0	0	0	0	0	0	0	0
DRC	KJV	5	0	0	0	0	6	2	0	1	0	2
KJV	WBT	1	0	0	0	0	0	0	0	0	0	0
WBT	ERV	7	0	1	0	0	1	1	0	0	0	0

Bible		lem alignments		lev alignments			other operations				error types		
source	target	correct	wrong	true pos	false pos	false neg	syn	hyper	hypo	co-hypo	WN	PP	AUX
MATT	GREAT	30	0	2	0	2	2	0	0	4	0	2	0
GREAT	GEN	53	0	0	0	3	2	0	0	2	0	2	0
GEN	RHE	30	0	1	0	0	8	0	2	2	0	0	0
RHE	DRC	2	0	0	0	0	0	0	0	0	0	0	0
DRC	KJV	4	0	0	0	0	6	2	0	2	0	0	0
KJV	WBT	1	0	0	0	0	0	0	0	0	0	0	0
WBT	ERV	4	0	1	0	0	0	0	0	0	0	0	0

LICENCE