ELECTRONIC TEXT REUSE ACQUISITION PROJECT

Detection of Text Reuse in Historical Texts

Marco Büchler, Greta Franzini, Maria Moritz, Emily Franzini, Gabriela Rotari

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

TABLE OF CONTENTS

1. Definition & motivation

2. Research on the characteristics

3. Characteristics: Qualitative research

4. Characteristics: Quantitative research

5. Research on the reuse process

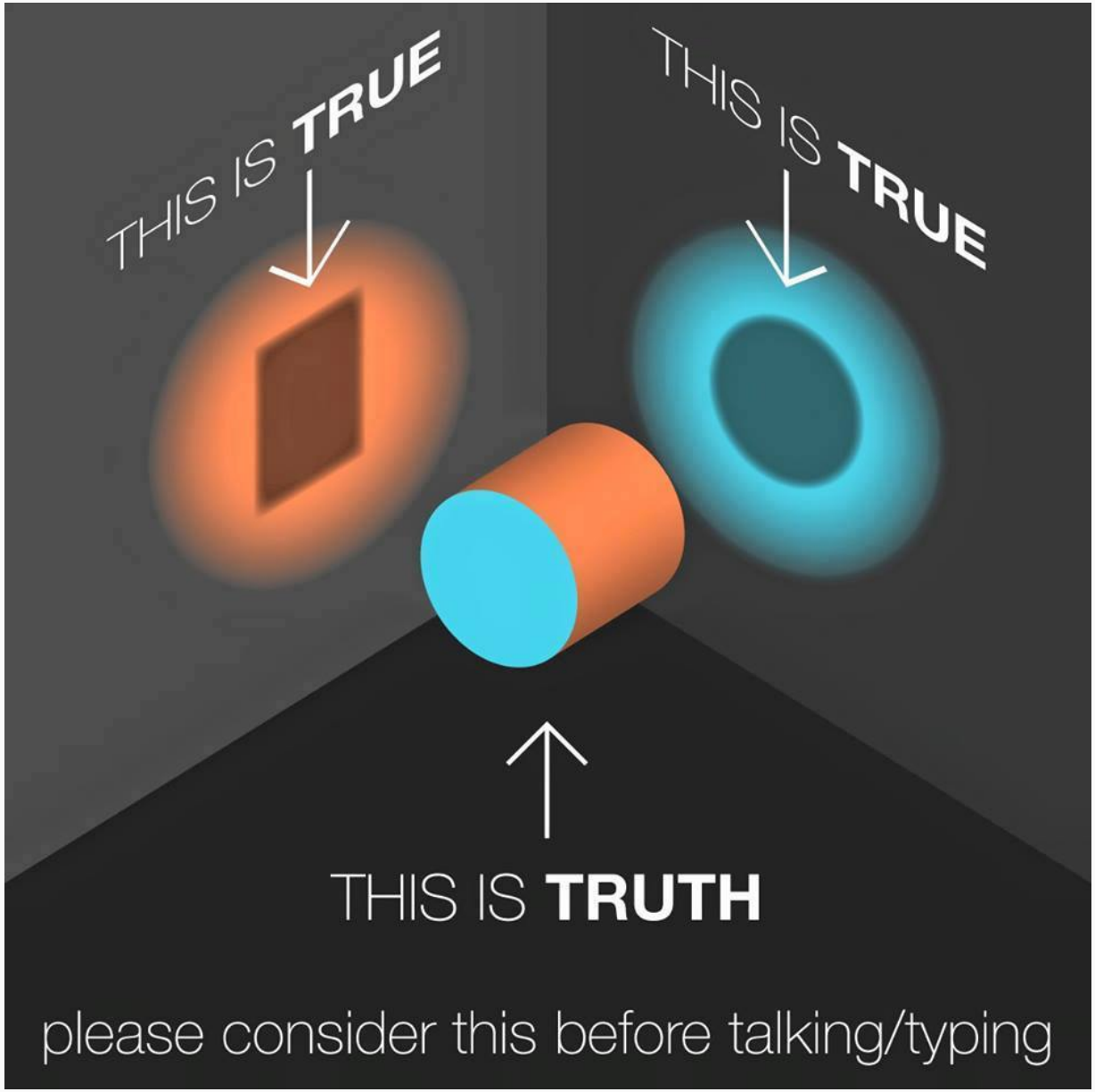
6. Process: Quantitative view

7. Process: Quantitative view

8. Results

DEFINITION & MOTIVATION

WHAT DO YOU ASSOCIATE WITH TEXT REUSE AND INTERTEXTUALITY?



TEXT REUSE

Text Reuse:

- spoken and written repetition of text across time and space.

For example:

- citations, allusions, translations.

Detection methods are needed to support scholarly work.

- E.g. they help to ensure clean libraries or identify fragmentary authors.

Text is often modified during the reuse process.

VENICE 2016 – TRACER TUTORIAL

WHO IS THIS PERSON?

“REUSE FROM SAME SOURCE”: COMMONALITIES & DIFFERENCES

WITTGENSTEIN'S "FAMILY RESEMBLANCE"

Family resemblance is an equivalence relation that clusters common objects of similar and not identical characteristics together.

Family resemblance is hierarchical such as in the examples before "Greta", "Franzinis", "Human", "creature".

FORENSIC VIEW

Evaluation of the reuse detection process by forensic criterions (standard in biometry):

- Universality: How univeral can a characteristic be? (example: for about 2% of all humans no fingerprint can be taken)
- Uniqueness: Different and independent "instances" should not share common characteristic.
- Permanence: How resistent is a characteristic over time?
- Collectability: Characteristics should be easy and simple to detect.
- Performance: It includes precision, speed and robustness of the measuring technique.
- Acceptability: Acceptance of the technique in (academic) usage.
- Circumvention: It should be as difficult as possible to cheat a detection system.

ETRAP'S OBJECTIVE

Title: eTRAP - electronic Text Reuse Acquisition Project

Premise: Language is a changing system. Compared to biometry the volatility is much higher.

- Research on the characteristics
  - What are good characteristics?
  - Which characteristics are stable and which are volatile and therefore not helpful in the detection process?
- Research on the reuse process
  - Begins with: Why do we quote what we quote?
  - Passes by: If changes in the reuse process happen, why do they happen and what is the model behind (if one exists)?
  - Ends with: Understanding paraphrases and allusions

Electronic Text Reuse Acquisition Project (eTRAP)

Interdisciplinary Early Career Research Group funded by the German Ministry of Education & Research (BMBF).

Budget: €1.6M.
Duration: March 2015 – February 2019. Research since October 2015.
Team: 4 core staff; 5-9 research & student assistants; Bachelor, Masters and PhD thesis students.

- Interdisciplinary: Classics, Computer Science, German Literature, Mathematics, Philosophy, Cognitive Psychology and Literature Studies.
- International: Currently from eight nationalities.

RESEARCH ON THE CHARACTERISTICS

TEXT REUSE CORE ELEMENTS

Motif: *"1. A minimal thematic unit"* (Prince, 2003, p. 55), set of core elements.

Core elements from an interdisciplinary standpoint:

- **Literature**: tracing MOTIFS
- **Cultural Studies**: tracing MEMES
- **Linguistics**: tracing PATTERNS
- **Computer Science**: tracing FEATURES
- **Forensics**: tracing MINUTIAE
- **Cognitive Psychology & Literature Studies**: tracing FIGURES OF MEMORY

CHARACTERISTICS: QUALITATIVE RESEARCH

KINDER- UND HAUSMäRCHEN

Seven editions of *Kinder- und Hausmärchen*: 1812-15, 1819, 1837, 1840, 1843, 1850, 1857.

Changes in:

- Size: from 156 to 201.
- Content: gruesome to mild.
- Style: Jacob scholarly, Wilhelm figurative.
- Language: Variants, diachronic evolution.

EXAMPLE CASE STUDY: SNOW WHITE

RQ: How to computationally detect a motif despite its variants?

For example:

- DE [Grimm][1]: ***Schneewittchen und die sieben Zwerge***
- EN [Briggs][2]: ***Snow White and the three robbers***
- IT [Calvino][3]: ***Bella Venezia e i dodici ladroni***
- SQ [von Hahn][4]: ***Schneewittchen und die vierzig Drachen***
- RU [Pushkin][5]: Сказка о мертвой царевне и о семи богатырях
- ...

A: We strike a balance between precision and recall. That is, finding the balance between a specific motif (Aarne-Thompson-Uther index) and its ontological root (Propp's typological unity).

HOW?

DATA COLLECTION AND CURATION

Tasks: Verify presence of motifs in different collections and record their "base form" as text reuse training data.

ISO Language Codes https://www.loc.gov/standards/iso639-2/php/code_list.php	GER						RUS		ITA	GLA		ARM	ENG			ARA		
Aarne-Thompson: 709	Grimm_1819 VIAF: 187449723	Grimm_1837 VIAF: 187449723	Grimm_1840 VIAF: 187449723	Grimm_1843 VIAF: 187449723	Grimm_1850 VIAF: 187449723	Grimm_1857 VIAF: 187449723	Pushkin_1833 VIAF: 312344013	Tsvetaeva_1911 VIAF: 185088476	Calvino_1956 VIAF: 181208131	Jacobs_1892 VIAF: 315397813	Bruford_1994 VIAF12471835	Hoogasian-Villa_1966 VIAF: 186329063	Campbell_1958 VIAF: 25969242	Taylor_1823 VIAF: 59071527	Briggs_1970 VIAF: 46803237	El-Shamy_1999 VIAF: 276573319	El Koudia_2003 VIAF: 5206198	Jason_1977 VIAF 9970253
D1300-D1379. Magic objects effect changes in persons																		
D1364. Object causes magic sleep	x	x	x	x	x	x	x	null	x	x	x	x	x	x	x	x	x	x
D1364.4. Fruit causes magic sleep	x	x	x	x	x	x	x	null	null	null	null	null	x	x	x	null	null	null
D1364.4.1. Apple causes magic sleep	x	x	x	x	x	x	x	null	null	null	null	null	x	x	x	null	null	null
D1364.9. Comb causes magic sleep	x	x	x	x	x	x	null	null	null	null	null	null	x	x	null	null	null	null
D1364.13. Cloth causes magic sleep	x	x	x	x	x	x	null	null	null	null	null	null	null	x	null	null	null	null
D1364.13.1. Lace causes magic sleep	x	x	x	x	x	x	null	null	null	null	null	null	null	x	null	null	null	null

Figure 1: Microsoft Excel matrix of motifs. Left column lists AT motifs in *Snow White* (AT 709); top row lists languages and collections covered.

Q400-Q599. Kinds of punishment	
Q411. Death as punishment	zu todt tanzen
Q414. Punishment: burning alive	glühende Pantoffeln, zu todt tanzen
Q414.4. Punishment: dancing to death in red-hot shoes	eiserne Pantoffeln, Feuer, glühend, anziehen, tanzen, Füße jämmerlich verbrannt, nicht aufhören, zu todt tanzen

Figure 2: Grimm motifs reduced to keywords.

DETECTION OF CROSS-LINGUAL MOTIFS

Train an (adapted) Named Entity Recognition (NER) tagger, ideally as language-independent as possible, to automatically annotate further fairy tales and texts.

EXAMPLE CASE STUDY: SNOW WHITE

RQ: How to computationally detect a motif despite its variants?

For example:

- DE [Grimm][1]: ***Schneewittchen und die sieben Zwerge***
- EN [Briggs][2]: ***Snow White and the three robbers***
- IT [Calvino][3]: ***Bella Venezia e i dodici ladroni***
- SQ [von Hahn][4]: ***Schneewittchen und die vierzig Drachen***
- RU [Pushkin][5]: Сказка о мертвой царевне и о семи богатырях
- ...

A: We strike a balance between precision and recall. That is, finding the balance between a specific motif (Aarne-Thompson-Uther index) and its ontological root (Propp's typological unity).

HOW?

THE NRC EMOTION LEXICON

The NRC (National Research Council Canada) **Emotion Lexicon:**

- The Roget Thesaurus
- 14,182 words types

Emotions: (Plutchik, 1980)	**Sentiments:**
anger	negative emotions
anticipation	positive emotions
disgust	
fear	
joy	
sadness	
surprise	
trust	

TAGGING EMPATHY

Classroom Questionnaires

- Empathy
- Identification
- Transportation

- Six- and ten-year-old children
- Y-Labor

- Data set

CHARACTERISTICS: QUANTITATIVE RESEARCH

TRACER: OVERVIEW

TRACER: suite of 700 algorithms developed by Marco Büchler.
Command line environment with no GUI.

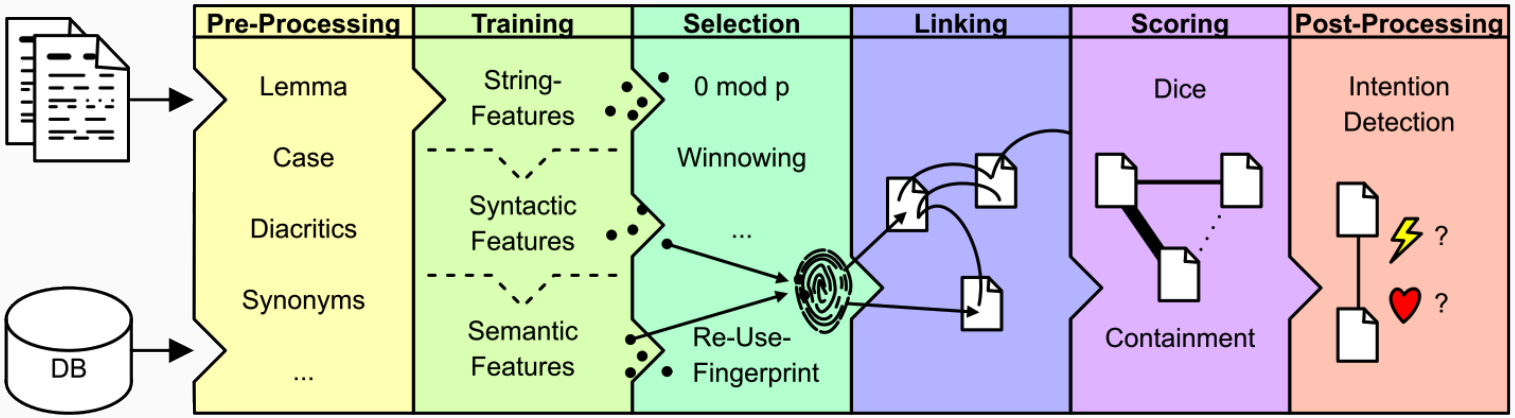


Figure 3: Detection task in six steps. More than 1M permutations of implementations of different levels are possible.

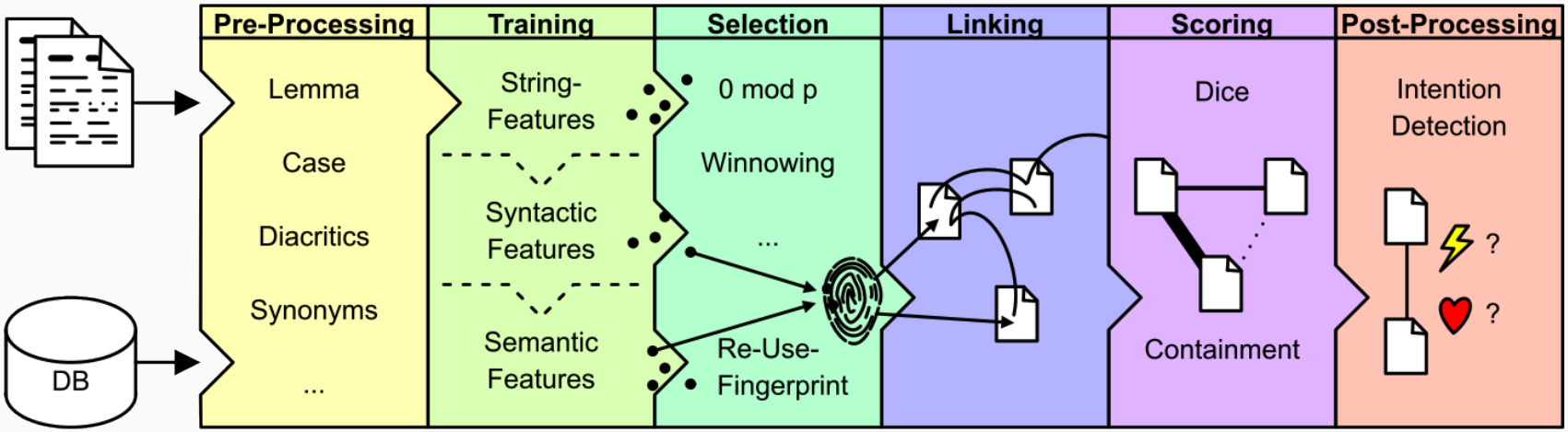
TRACER tested on: Ancient Greek, Arabic, Coptic, English, German, Hebrew, Latin, Tibetan.

TRACER: DISSEMINATION

Webpage: `http://www.etrap.eu/research/tracer`
Repository: `http://vcs.etrap.eu/tracer-framework/tracer.git`
Upcoming tutorials:

- **AIUCD 2017** (Jan 2017): pre-conference workshop with DiXiT, Rome, Italy.
- **DATeCH 2017** (May 2017): pre-conference workshop, Göttingen, Germany.
- Three more tutorials in 2017 pending confirmation.

MOTIVATION FOR AN ANALYSIS OF CORE COMPONENTS



Analysing core component affects the levels Pre-processing, Training/ Featuring and Selection.

SETUP & FACT FILE

- Two lists of Biblical and Medieval German idioms each
- Idioms as they are widely spread
- 25 participants have been asked to remove those words so that they can still identify the idiom
- Result data-set: 10,000 datasets by 2x200 idioms (Biblical and Medieval) with 25 participants each
- Objective: 25 participants/interraters enable research on the human process of feature selection: What do humans select as relevant?
- Data-set will be made publicly available by 01/2017.

REMOVED WORDS

- **Bibel**: *ein* (563), *die* (276), *das* (193), *sein* (176), *den* (170), *der* (169), *wie* (131), *und* (127), *im* (107), *ist* (105), *etwas* (94), *einen* (93), *in* (92), *eine* (88), *auf* (78), *sich* (76), *sein* (73), *jemanden* (71), *haben* (58), *!* (55), *,* (50), *von* (46), *vom* (43), *jemandem* (42), *gehen* (41), *das* (38), *machen* (38), *werden* (38), *dem* (37), *mit* (37)
- **Mittelalter**: *ein* (563), *die* (276), *das* (193), *sein* (186), *einen* (172), *ein* (140), *und* (117), *sich* (111), *haben* (107), *auf* (98), *dem* (93), *!* (85), *der* (77), *,* (75), *eine* (64), *mit* (64), *jemandem* (59), *jemanden* (46), *in* (40), *ins* (40), *am* (38), *kommen* (37), *einer* (35), *machen* (35), *wie* (34), *aus* (33), *es* (31), *das* (30), *legen* (29)

RESULTS OF PARTICIPANTS

Average feature densities $\mathcal{F}^{\mathcal{B}} = 0.7585$ und $\mathcal{F}^{\mathcal{M}} = 0.7699$ form baseline.

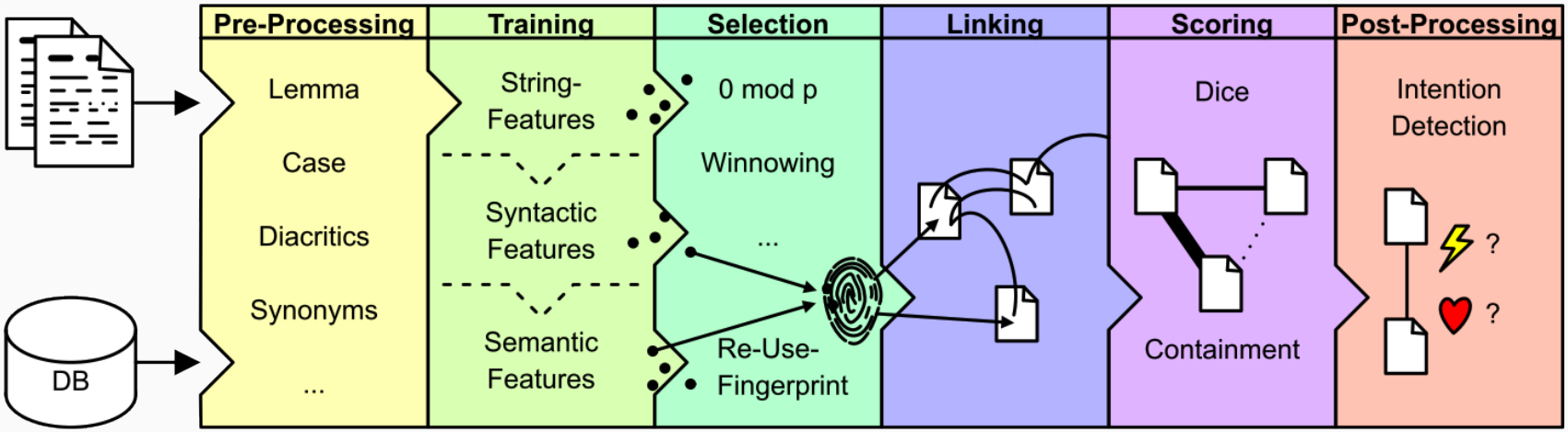
Part of Speech-Tag	Wortartklasse
n	noun
v	verb
t	participle
a	adjective
d	adverb
l	article
g	particle
c	conjunction
r	preposition
p	pronoun
m	numeral
i	interjection
e	exclamation
u	punctuation

	n	v	t	a	d	l	g	c	r	p	m	u
Bibel	0.98	0.86	0.81	0.95	0.69	0.39	0.71	0.70	0.72	0.56	0.80	0.58
Mittelalter	0.98	0.88	0.93	0.95	0.79	0.42	0.81	0.71	0.79	0.49	0.84	0.52

RESULTS OF PARTICIPANTS

- Inter-PoS analysis (dependencies between PoS tags)
- Interrater analysis
- Making data-sets available (including tagged data)

MOTIVATION FOR AN ANALYSIS OF CORE COMPONENTS



Analysing core component affects the levels Pre-processing, Training/ Featuring and Selection.

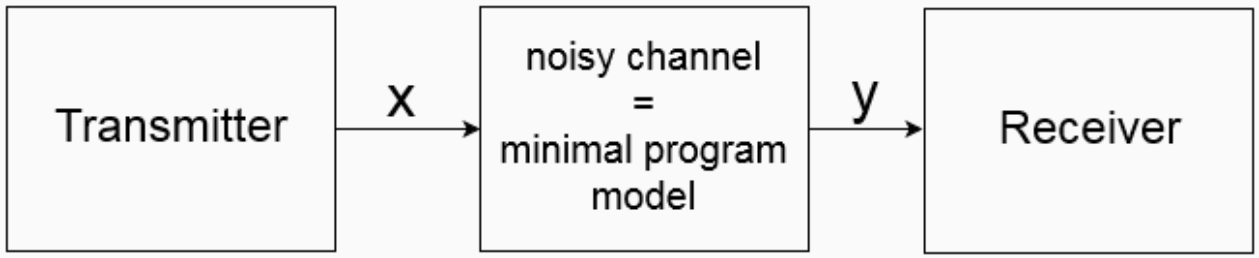
RESEARCH ON THE REUSE PROCESS

Paraphrasing and non-literal reuse challenges many approaches:

- Alzahrani et al. (2012)
  - study n-gram-, syntax-, and semantic-based detection approaches;
  - they find: as soon as reuse is slightly modified (words changed) most approaches fail.
- Barrón-Cedeño et al. (2013)
  - experiment with paraphrasing to improve plagiarism detection;
  - they found that complex paraphrasing with a high density challenges plagiarism detection, and
  - that lexical substitution is the most frequent plagiarism technique.

APPROACH

- Inspired by
  - Shannon's noisy-channel: for a given degree of noise, it is possible to transmit digital data error-freely up to a computable maximum rate in a communication channel (Shannon, 1949),
  - Kolmogorov Complexity: describes the length of the shortest program that produces an output string (Li and Vitáni, 2008),
  - Generative Story (similar to IBM's alignment model) (e.g., Shannon, 1948),
- we study Ancient text reuse to understand how text is transferred.
  - Identify operations to characterize morphological & semantic changes
  - Design an algorithm which applies these OPs to our datasets
  - Transform one text excerpt into another by a minimum OP set



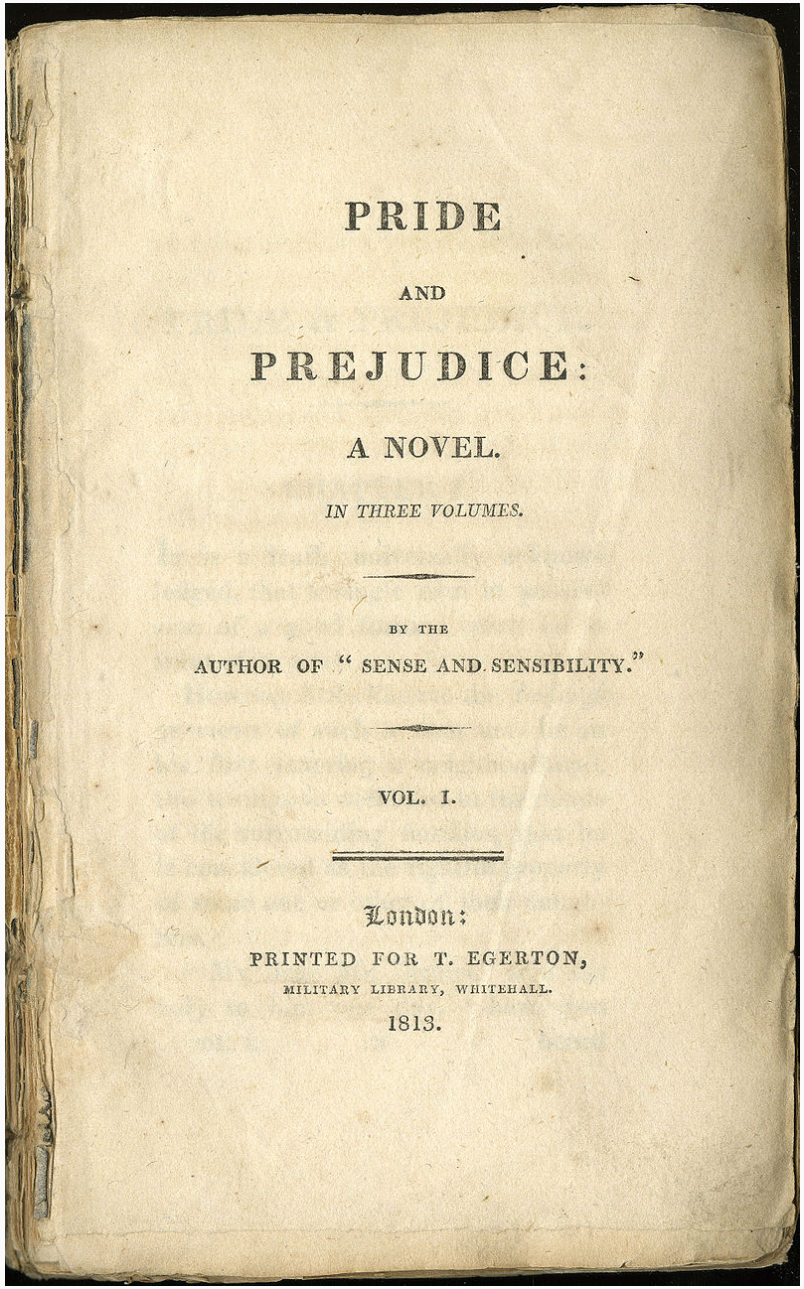
TRANFORMATION OPERATIONS

Table 1: Operation list for the automated approach

operation	description	example
NOP(reuse_word, orig_word)	Original and reuse word are equal.	*NOP(maledictus,maledictus)*
upper(reuse_word, orig_word)	Word is lowercase in reuse and uppercase in original.	*upper(kai,Kai)* - in Greek
lower(reuse_word, orig_word)	Word is uppercase in reuse and lowercase in original.	*lower(Gloriam,gloriam)*
lem(reuse_word, orig_word)	Lemmatization leads to equality of reuse and original.	*lem(penetrat,penetrabit)*
repl_syn(reuse_word, orig_word)	Reuse word replaced with a synonym to match original word.	*repl_syn(magnificavit,glorificavit)*
repl_hyper(reuse_word, orig_word)	Word in Bible verse is a hyperonym of the reused word.	*hyper(cupit,habens)*
repl_hypo(reuse_word, orig_word)	Word in Bible verse is a hyponym of the reused word.	*hypo(dederit,tollet)*
repl_co-hypo(reuse_word, orig_word)	Reused word and original have the same hyperonym.	*repl_co-hypo(magnificavit,fecit)*
NOPmorph(reuse_tags, orig_tags)	Case or PoS did not change between reused and original word.	*NOPmorph(na,na)*
repl_pos(reuse_tag, orig_tag)	Reuse and original contain the same cognate, but PoS changed.	*repl_pos(n,a)*
repl_case(reuse_tag, orig_tag)	Reuse and original have the same cognate, but the case changed.	*repl_case(g,d)* - cases genitive, dative
lemma_missing(reuse_word, orig_word)	Lemma unknown for reuse or original word.	*lemma_missing(tentari, inlectus)*
no_rel_found(reuse_wword, orig_word)	Relation for reuse or original word not found in AGWN.	*no_rel_found(gloria,arguitur)*

PROCESS: QUANTITATIVE VIEW

JANE AUSTEN'S PRIDE & PREJUDICE



PRIDE

AND

PREJUDICE:

A NOVEL.

IN THREE VOLUMES.

BY THE

AUTHOR OF "SENSE AND SENSIBILITY."

VOL. I.

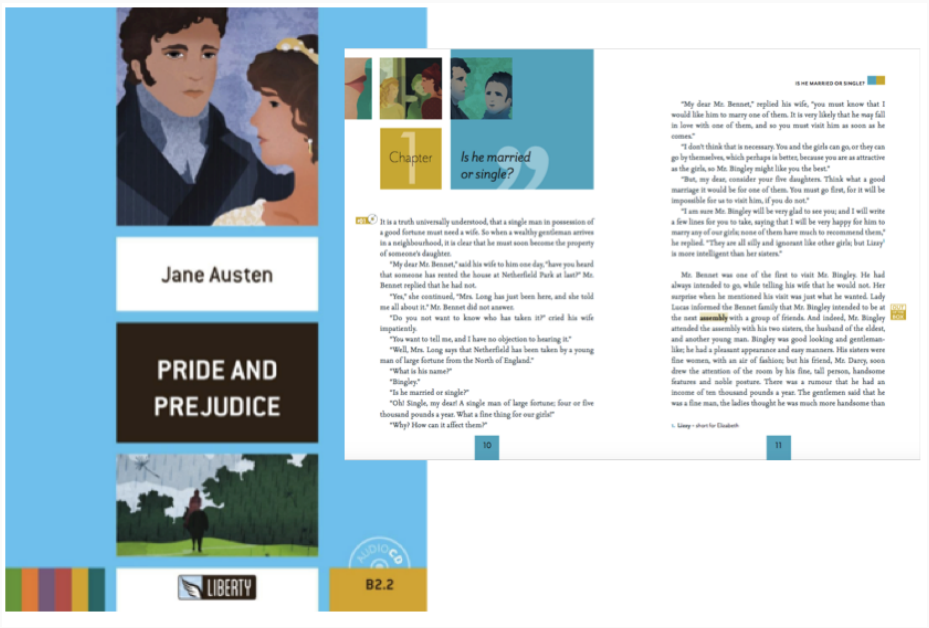
London:

PRINTED FOR T. EGERTON,

MILITARY LIBRARY, WHITEHALL.

1813.

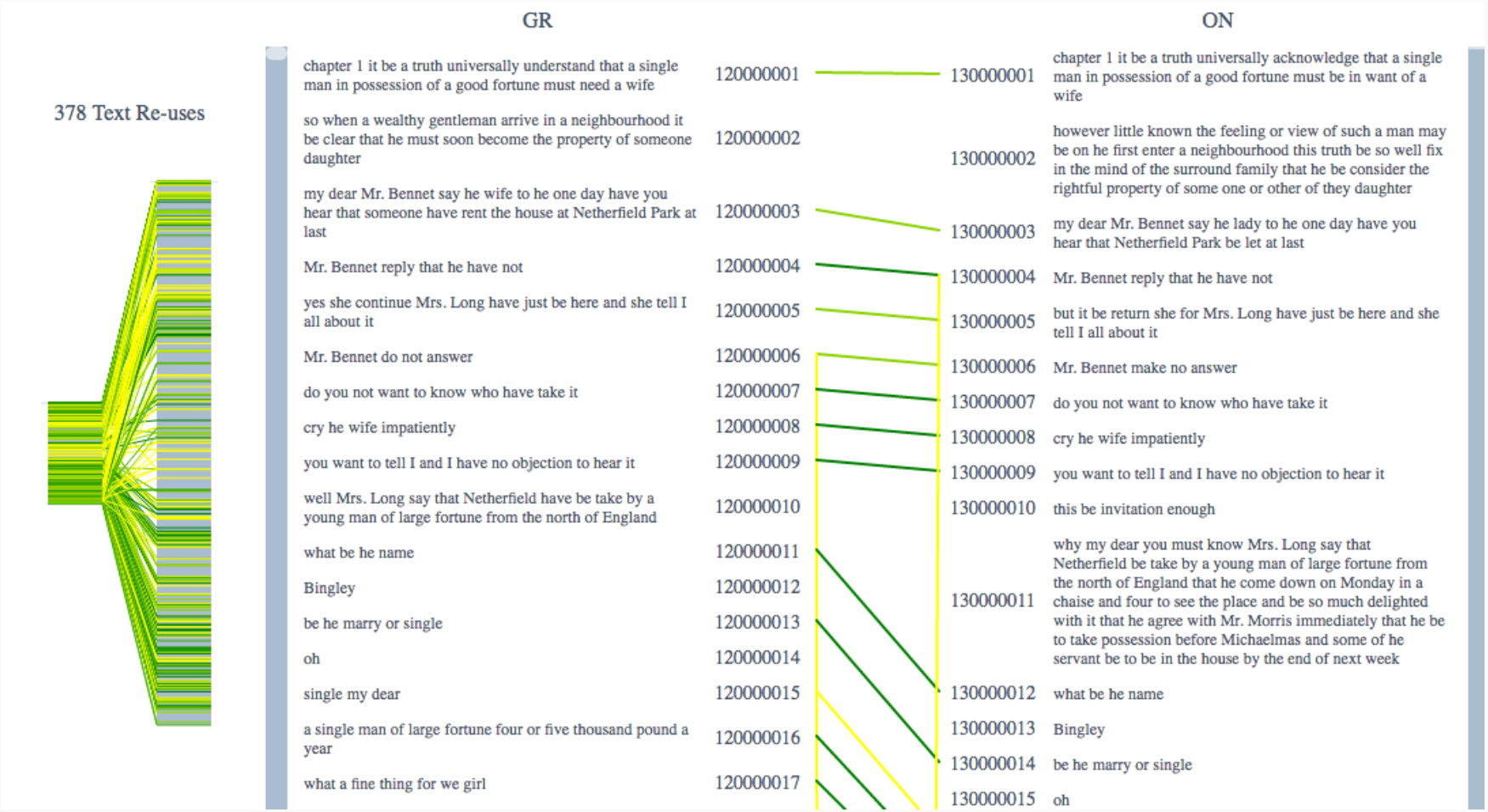
GRADED READER



Definition:

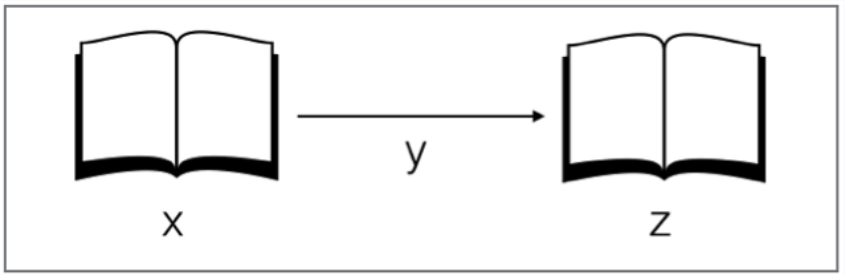
Graded readers are “simplified books written at varying levels of difficulty for second language learners", which “cover a huge range of genres ranging from adaptation of classic works of literature to original stories, to factual materials such as biographies, reports and so on” [Waring 2012].

AUTOMATIC ALIGNMENT OF ORIGINAL NOVEL WITH GRADED READER



To computationally analyse the process Y and classifying the changes:

- Do the changes follow strict rules?
- Do they form patterns?
- Can they be computationally reproduced?



Categories of changes:

- Cognitive
- Structural
- Cognitive and structural

TESTING THE SIMPLIFICATION WITH READABILITY TESTS

Readability tests aim to classify texts by their degree of complexity and understandability. Measured primitives are sentence length and difficulty of the words.

Two tests, the ARI score and the Dale-Chall-Index have been selected:

The ARI score is based on the word length and the sentence length:

$$R_{ARI} = 4.71\left(\frac{characters}{words}\right) + 0.5\left(\frac{words}{sentences}\right) - 21.43 \quad (1)$$

The Dale-Chall-Index is based on the word frequency (3000 most frequent words) and the sentence length:

$$R_{DCI} = 0.1579\left(\frac{difficult\ words}{words} * 100\right) + 0.0496\left(\frac{words}{sentences}\right) \quad (2)$$

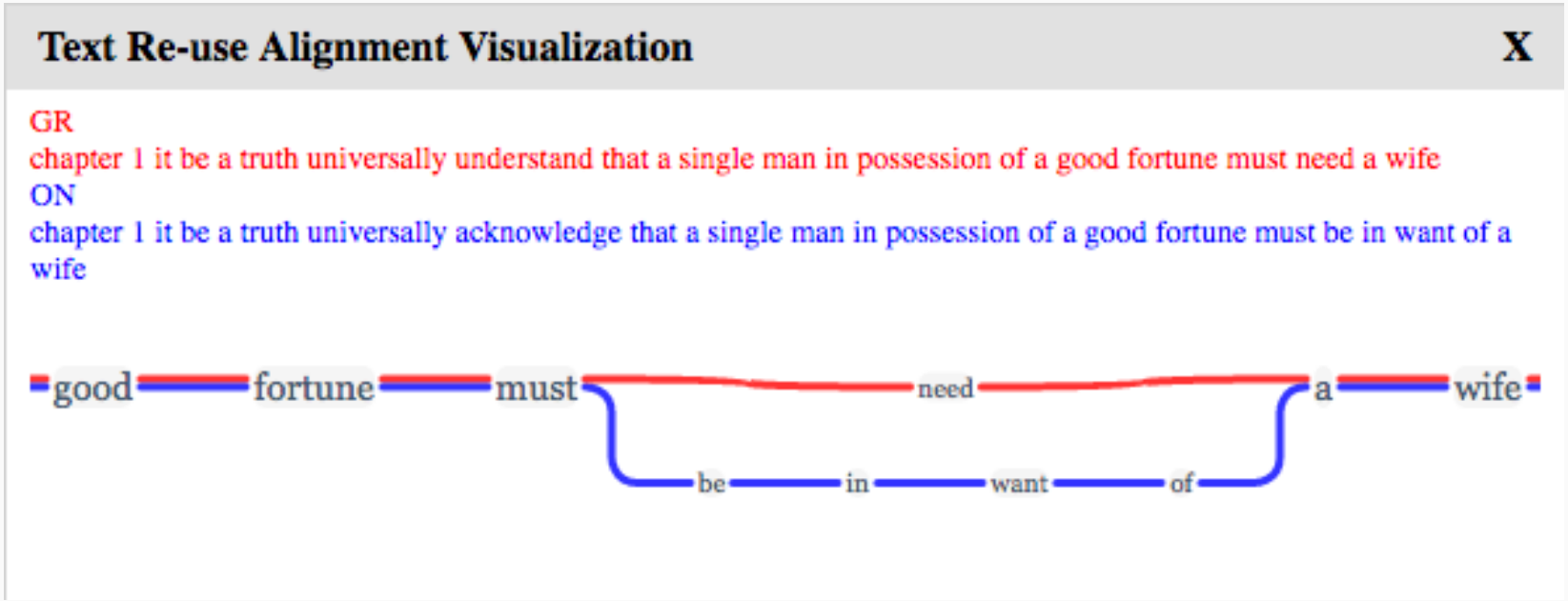
RESULTS OF THE SIMPLIFICATION WITH READABILITY TESTS

Readability test result matrix:

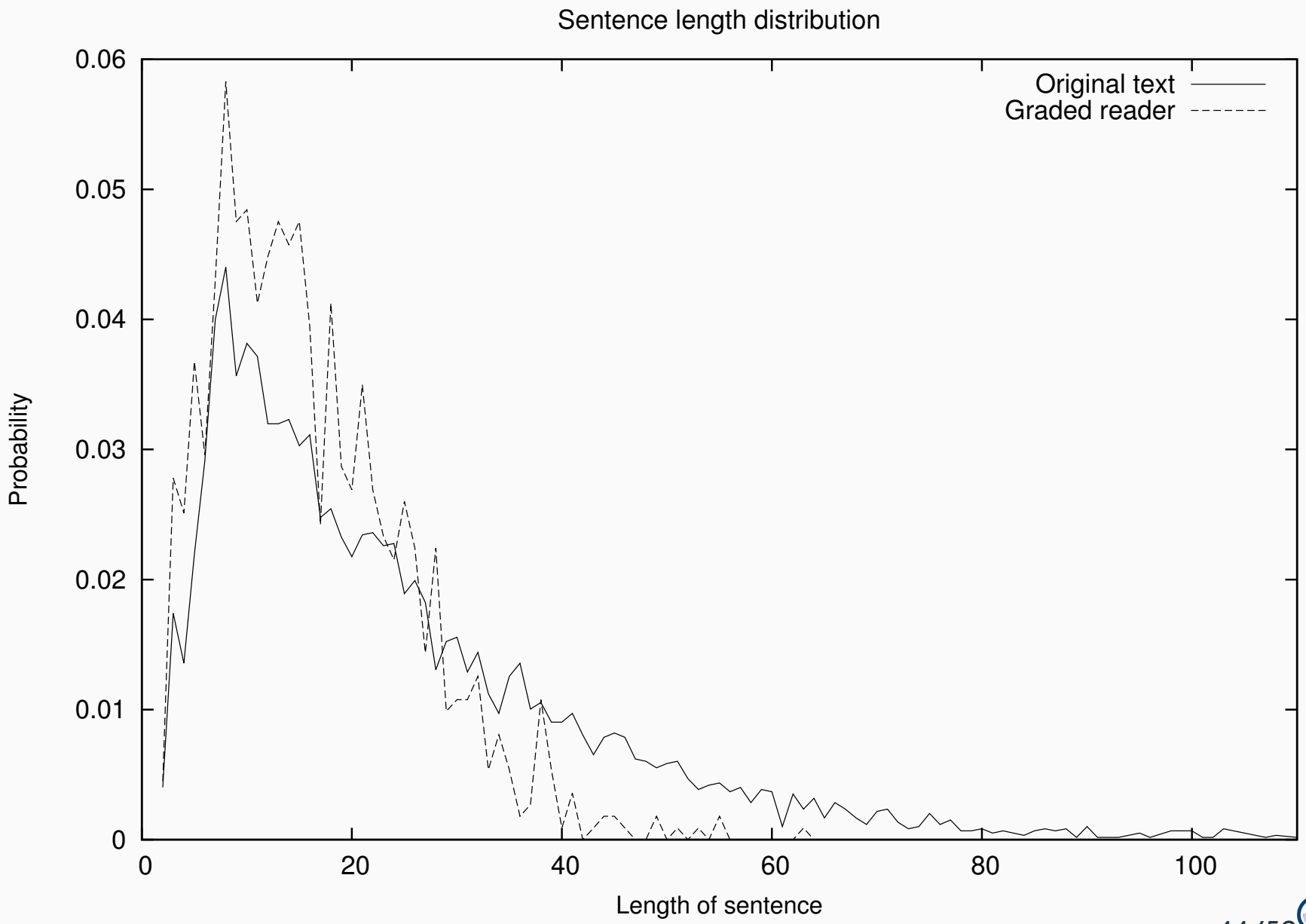
	ARI	Dale-Chall
Original Novel	14-15 year olds	14-16 year olds
Graded Reader	11-12 year olds	11-13 year olds

SIMPLIFICATION & SENTENCE LENGTH

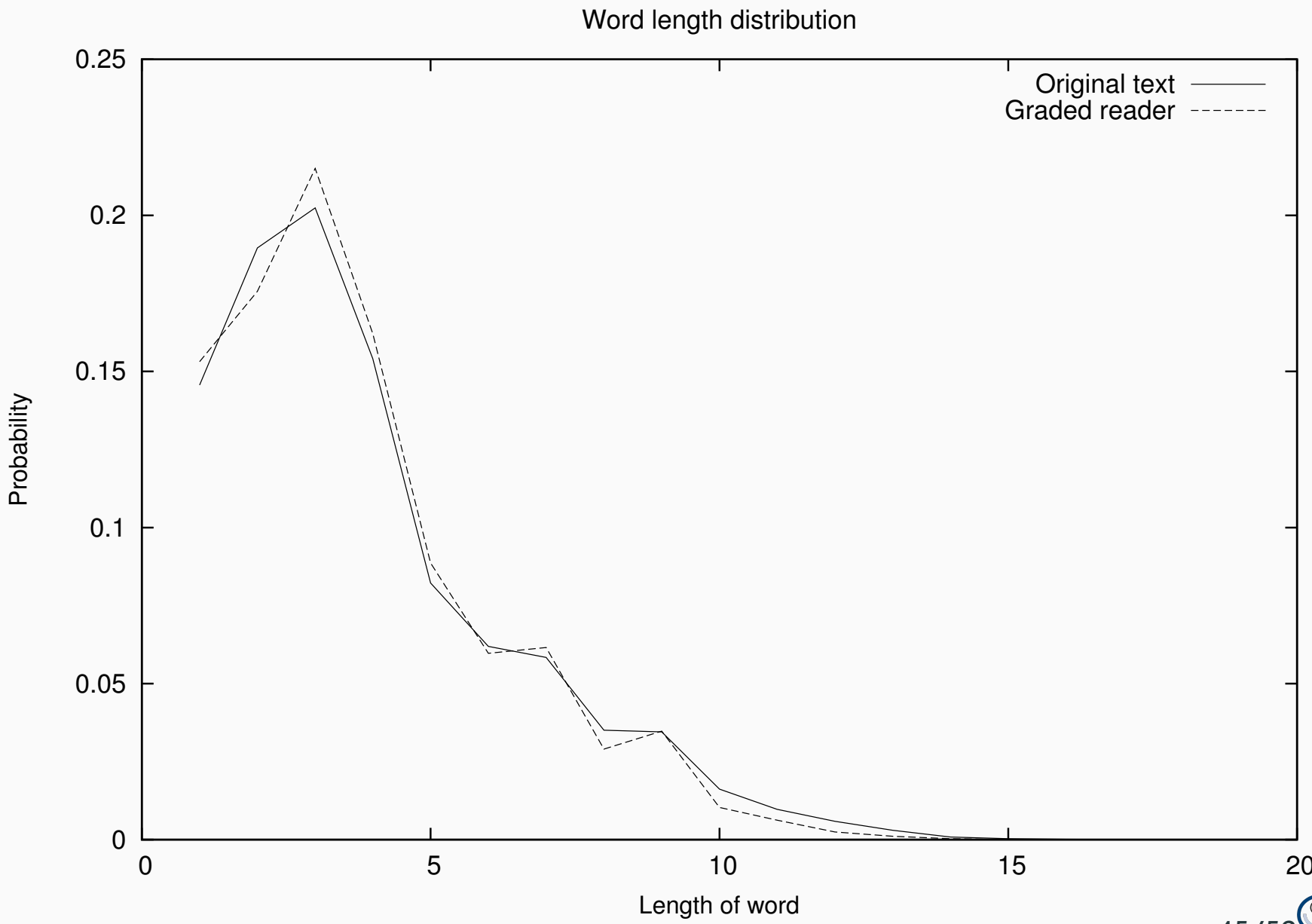
An example of a structural text simplification > many-to-one.



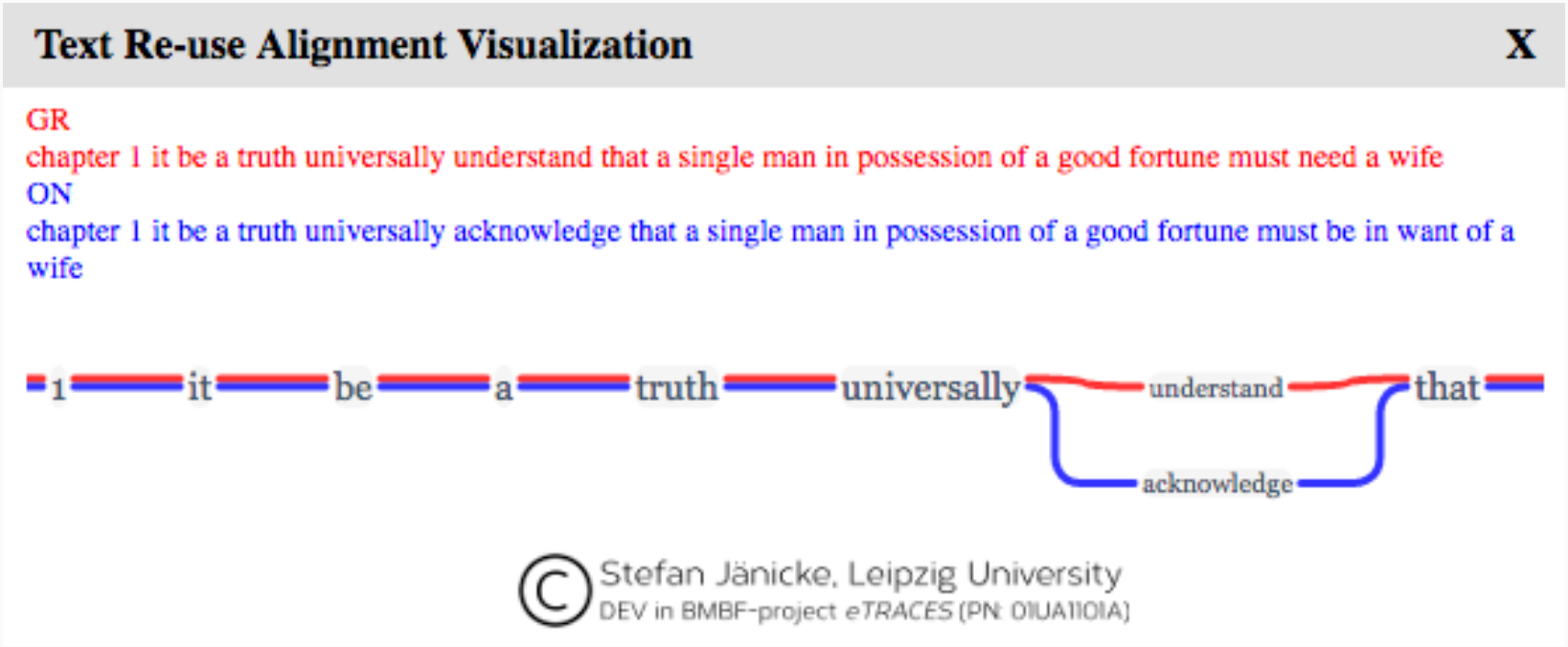
COMPARISON OF SENTENCE LENGTH



COMPARISON OF WORD LENGTH



EXAMPLE OF WORD REPLACEMENT



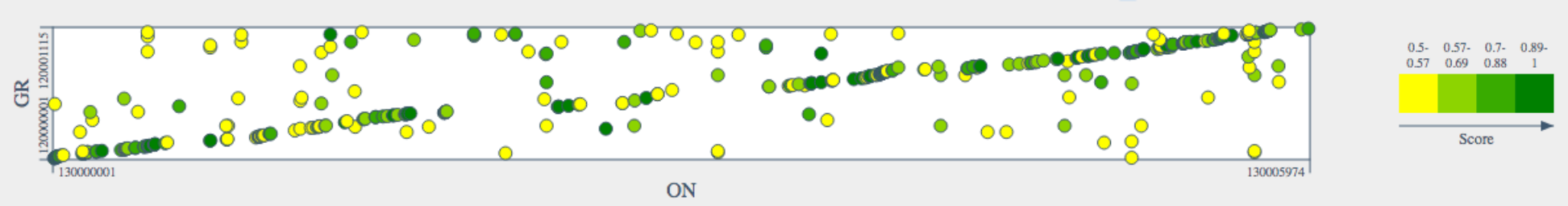
Conclusion: The simplification of words is provided by using easier and more frequent words instead of shortened words.

DIFFERENCE ANALYSIS: WORDS APPEARING ONLY IN THE ORIGINAL

Word	Frequency	Word	Frequency
upon	75	table	31
least	65	astonishment	30
acquaintance	63	fancy	30
either	59	attempt	29
whose	59	dine	29
dare	53	beg	28
regard	53	depend	28
determine	47	highly	28
scarcely	45	satisfaction	28
ladyship	42	acknowledge	27
former	38	credit	27
put	36	thus	27
amiable	35	disposition	26
deal	34	exceedingly	26
design	32	praise	26

MACRO SCALE: VISUALISATION OF THE SELECTION PROCESS

The Dotplot view of original novel against the graded reader on a sentence-wise segmentation uncovers which passages were taken over in the graded reader and which not:



PROCESS: QUANTITATIVE VIEW

DATA-SETS - ANCIENT GREEK AND LATIN DATA-SET

"Salvation for the Rich"

Clement of Alexandria

Christian theologian, 2nd cent.

- Known for his retelling of biblical excerpts
- Reuse annotated by Biblindex team (Mellerin, 2014; Mellerin, 2016)
- We obtain 199 verse-reuse-pairs
- Pointing to 15 Bible books

Extracts from 12 works & 2 collections

Bernard of Clairvaux

French abbot, 12th cent.

- Known for his influence on the Cistercian order and his work in biblical studies
- Reuse extracted by Biblindex team (Mellerin, 2014; Mellerin, 2016)
- We obtain 162 verse-reuse-pairs
- Pointing to 31 Bible books

The data was tokenized and punctuation was kept but ignored in the analyses.

BIBLICAL REUSE EXAMPLES

more literal	Bible verse	Bernard reuse
Proverbs 18 3	**impius cum in profundum venerit** peccatorum **contemnit** sed sequitur eum ignominia et obprobrium (*When the wicked man is come into the depth of sins, also contempt comes but ignominy and reproach follow him)*	**Impius** , **cum venerit in profundum** malorum , **contemnit** (*When the wicked man is come into the depth of evil*)

less literal	Bible verse	Clement reuse
1Cor 13 13	νυνὶ **δὲ μένει πίστις** , **ἐλπίς** , **ἀγάπη** , **τὰ τρία ταῦτα μείζων δὲ τούτων ἡ ἀγάπη** (*And now remain faith, hope, love, these three; but the greatest of those is love.*)	**πίστει** καὶ **ἐλπίδι** καὶ **ἀγάπῃ** (*faith, and hope, and love* - in dative case)
		ἀγάπην , **πίστιν** , **ἐλπίδα** (*love, faith, hope* - in accusative case)
		μένει δὲ τὰ τρία ταῦτα , **πίστις** , **ἐλπίς** , **ἀγάπη** · **μείζων δὲ** ἐν **τούτοις ἡ ἀγάπη** *(and remain these three, faith, hope, love; but the greatest among them is love)*

non-literal	Bible verse	Clement reuse
Mt 12 35	ὁ ἀγαθὸς ἄνθρωπος ἐκ τοῦ ἀγαθοῦ θησαυροῦ ἐκβάλλει ἀγαθά , καὶ ὁ **πονηρὸς** ἄνθρωπος ἐκ τοῦ πονηροῦ θησαυροῦ ἐκβάλλει **πονηρά** . (*A good man out of good storage brings out good things , and an evil man out of the evil storage brings evil things .*)	Ψυχῆς , τὰ δὲ ἐκτός , κἂν μὲν ἡ ψυχὴ χρῆται καλῶς , καλὰ καὶ ταῦτα δοκεῖ , ἐὰν δὲ **πονηρῶς** , **πονηρά** , ὁ κελεύων ἀπαλλοτριοῦν τὰ ὑπάρχοντα (*[are whitin the] soul, and some are out, and if the soul uses them good, those things are also thought of as good, but if [they are used as] bad, [they are thought of as] bad; he who commands the renouncement of possessions*)

LINGUISTIC SUPPORT – LEMMA RESOURCES

We aggregate:

- Biblindex' Lemma Lists
  - 65,537 Biblical Greek entries
  - 315,021 Latin entries
- Classical Language Tool Kit (CLTK) (Johnson et al., 2014)
  - 953,907 Ancient Greek words
  - 270,228 Latin words
- Greek New Testament of the Society of Biblical Literature[1] & Septuaginta (Rahlfs, 1935a; UPenn) 59,510 word-lemma-pairs

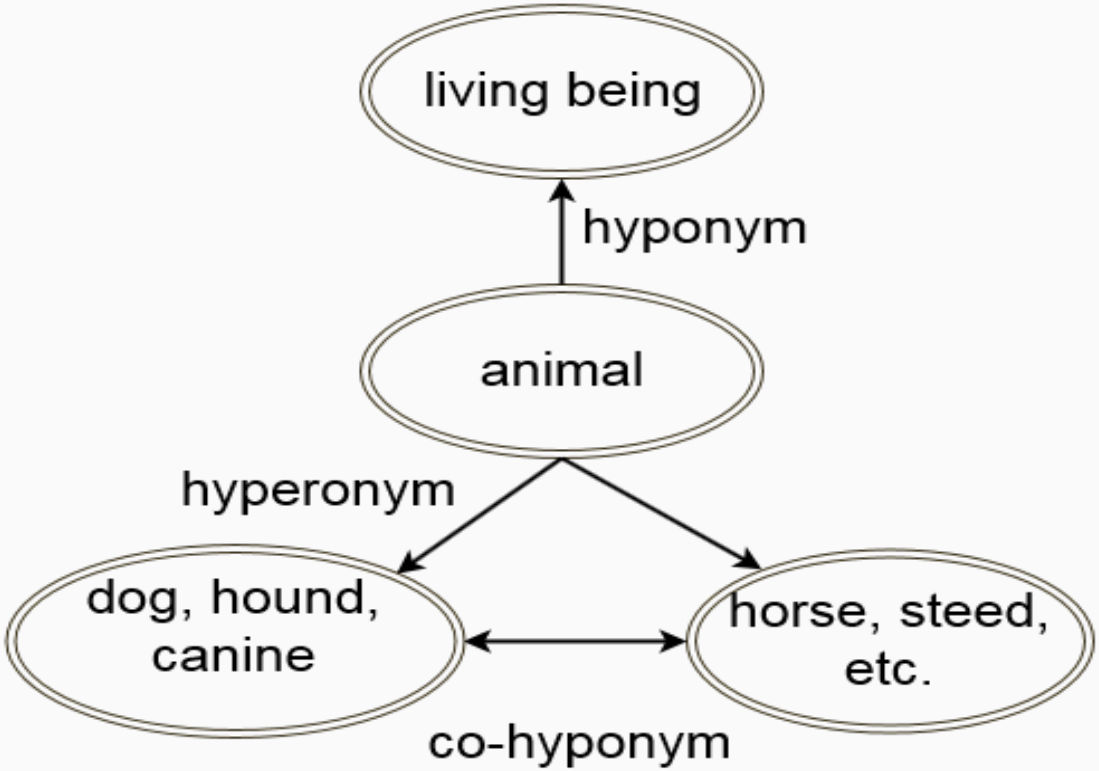
[1] Logos Bible Software `http://sblgnt.com/about/`

LINGUISTIC SUPPORT – ANCIENT GREEK WORDNET (AGWN)

99K synsets

of which 33K contain Ancient Greek and 27K Latin words

(Bizzoni et al., 2014; Minozzi, 2009)



RESULTS

LITERAL SHARE OF THE REUSE (RQ1)

What is the extent of non-literal reuse in our datasets?

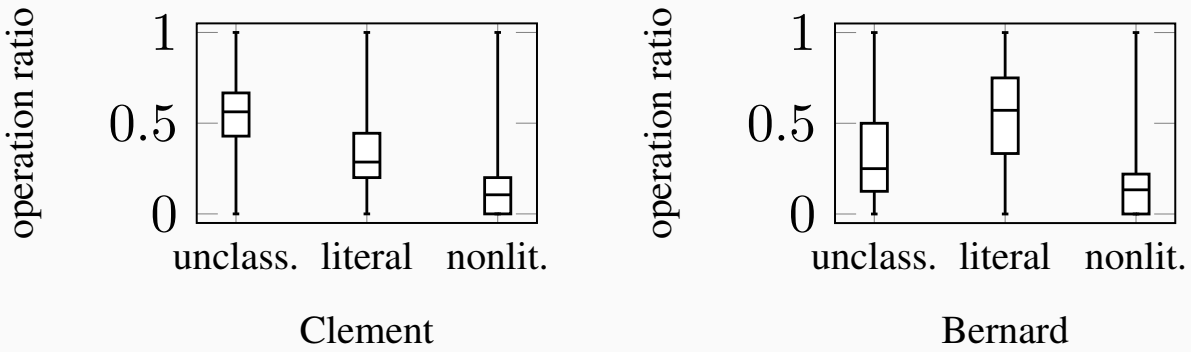


Figure 4: Ratios of operations in reuse instances. literal: NOP, lem, lower, etc.; nonlit: syn, hyper, etc.

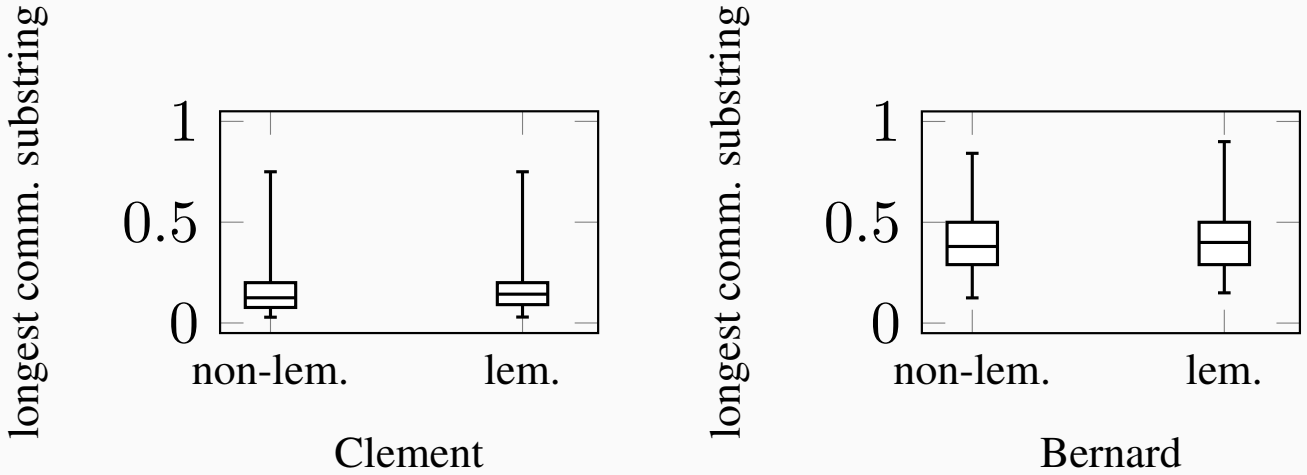


Figure 5: Ratios of literal overlap between reuse instances and originals.

AUTOMATED APPROACH (RQ2.1)

How is the non-literally reused text modified in our datasets? (RQ2)
How can linguistic resources support the discovery of non-literal reuse? (RQ2.1)

Table 2: Absolute numbers of operations identified automatically.

	literal				non-literal				unclassified		
	NOP	upper	lower	lem	syn	hyper	hypo	co-hypo	no_rel_found	lem_missing	total
Greek	337	6	0	356	153	20	14	101	563	639	2189
Latin	587	0	44	102	60	14	28	68	347	85	1335

AUTOMATED APPROACH (RQ2.1) - COVERAGE VALUES

Operations that successfully looked up a lemma: `lem_success={lem, syn, repl_hyper, repl_hypo, repl_co-hypo, no_rel_found}`, with `lem_missing` representing not found tokens in the lemmata.

$$\text{COV}_{\text{lem}} = \frac{\sum_{\text{Occ}(o)} o \in \text{lem\_success}}{\sum_{\text{Occ}(o)} o \in \text{lem\_success} \cup \{\text{lem\_missing}\}}$$

$$\text{COV}_{\text{AGWN}} = \frac{\sum_{\text{Occ}(o)} o \in \text{agwn\_success}}{\sum_{\text{Occ}(o)} o \in \text{agwn\_success} \cup \{\text{no\_rel\_found}\}}$$

We obtain a cov_{lem} of **0.65** for our Greek and **0.88** for the Latin data-set. And a cov_{AGWN} of **0.34** for our Greek and **0.33** for our Latin data-set.

Language resources help to get an idea of reuse components.

CONTACT

Visit us


http://www.etrap.eu


contact@etrap.eu

Stealing from one is plagiarism, stealing from many is research
(Wilson Mitzner, 1876-1933)

SPONSORED BY THE

LICENCE