Scaling Historical Text Re-use
eTRAP Research Group

2nd Workshop on Big Humanities Data at IEEE Big Data 2014

Marco Büchler[1], Greta Franzini[2], Emily Franzini[2], Maria Moritz[2]

[1] Göttingen Center for Digital Humanities, University of Göttingen
[2] Chair for Digital Humanities, Leipzig University

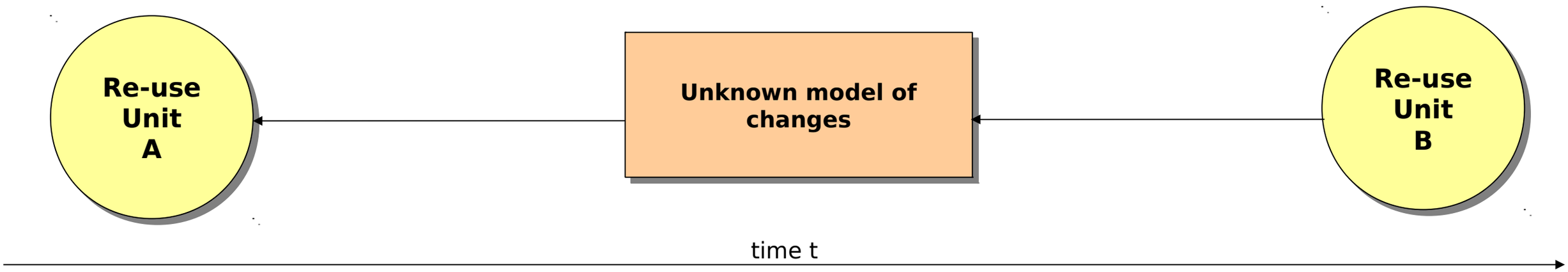
October 27th 2014

What is (Historical) *Text Re-use*?

General: Text Re-use describes the spoken and written repetition of content.

Example: quotations, paraphrases but also translations

Historical changes: language evolution, different dialects, “spelling errors” but also copy errors (by monks in the Mid-ages)



Historical Text Re-use as an Opportunity for Computer Science and Humanities

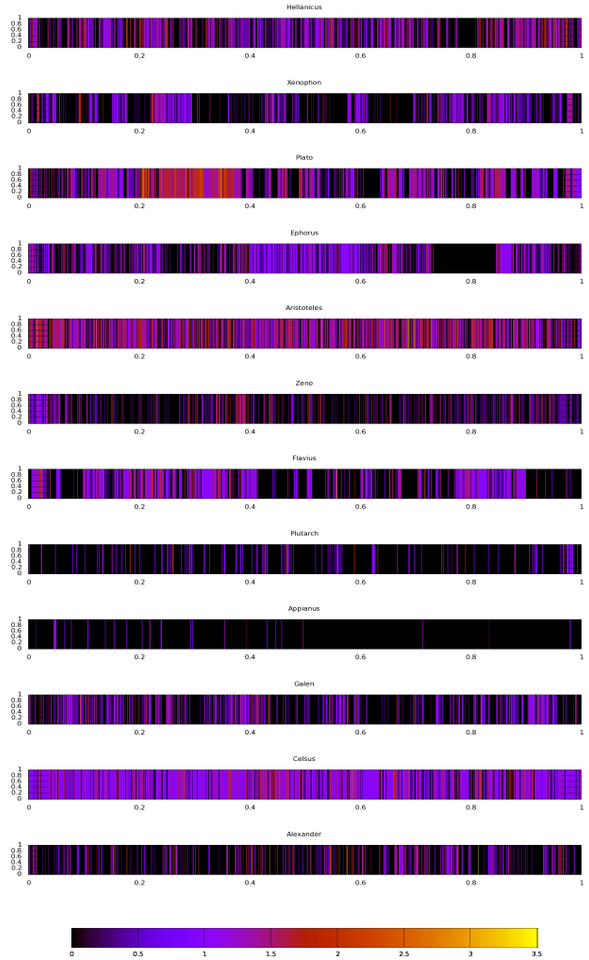
---

Question: Why is Text Re-use so fundamental for Humanities and Computer Science?

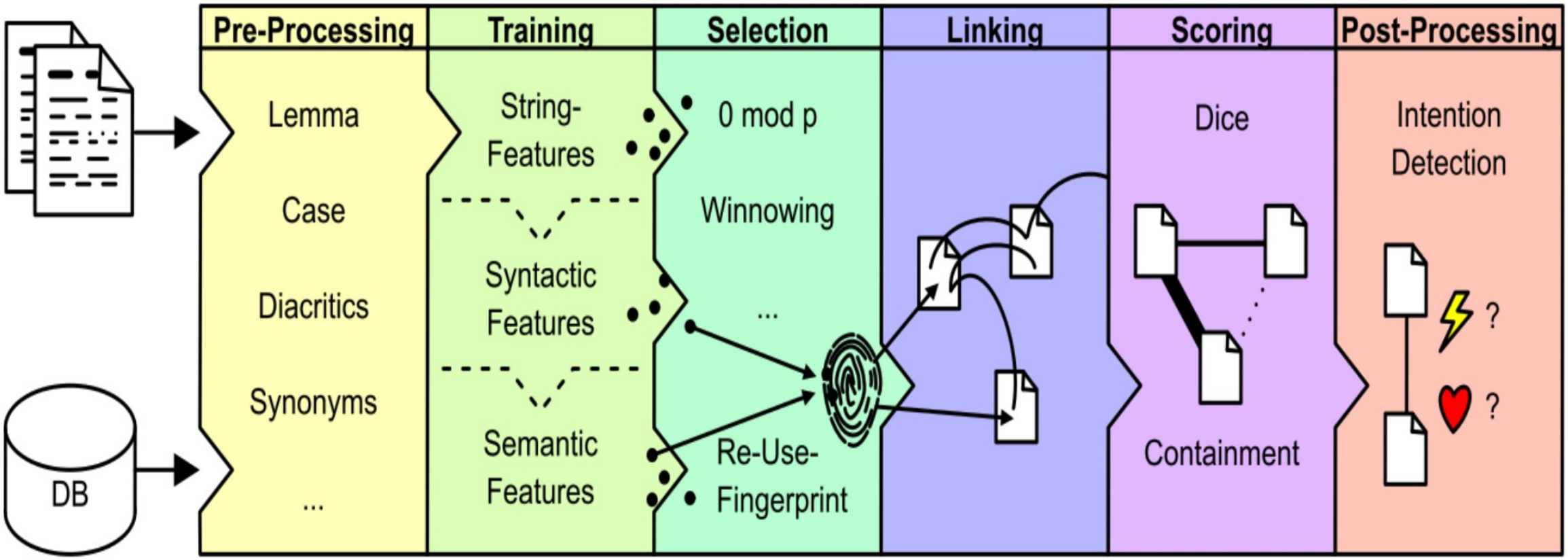
Premise: the amount of digitally available data grows exponentially (Big Data)

Humanities
- Lines of transmissions and textual criticism
- Transmissions of ideas/thoughts under different circumstances and conditions

Computer Science:
- Text Decontamination for stylometry and authorship attribution, dating of texts
- gen. Text Mining, Corpus Linguistic



Approach



Implemented in TRACER software: more than one million permutations of implementations of different levels are recently possible

Text Re-use on English Bible versions

Why does the use of the Bible make sense?:

- The Bible is **easy to evaluate**.

- There are different editions written for **different purposes**.

Text Re-use on English Bible versions **Evaluation**

Example: book Genesis, chapter 1, verse 1

ASV	In the beginning God created the heavens and the earth.
BBE	At the first God made the heaven and the earth.
DBY	In the beginning God created the heavens and the earth.
KJV	In the beginning God created the heaven and the earth.
Webster	In the beginning God created the heaven and the earth.
WEB	In the beginning God created the heavens and the earth.
YLT	In the beginning of God‘s preparing the heavens and the earth.

Reduced Bibles: all seven reduced Bible versions contain “only” the 28632 verses contained in all seven editions.

8 forensic aspects to algorithms (Jain 2005, Maltoni 2009)

- Acceptability

- Circumvention

- Collectability

- **Performance**

- Permanence

- Selection

- Distinctiveness

- Universality

Why does performance matter?

A theoretical experiment:

- **Assuming** 1 million books, each with 50 pages, each page has 20 sentences; this brings 1 billion sentences

- **Brute Force Linking:** Comparing 1 billion (10^9) sentences pairwise with each other; result 10^{18} pairwise comparisons

- **Assuming** a throughput of 1000 comparisons/sec

- **Result**: 10^{15} seconds or 31.7 million years of runtime single-threaded

1st step: Feature based linking

1: Linking by parametrisation through Feature Density

LINKING ANALYSIS DEPENDING ON *Feature Density* $\mathcal{F}$ FOR UNIQUE LINKS (UL), LINKED LINKS (LL), AVERAGE LINKS (AL) AND THE BRUTE FORCE SCORE (BFS). *UL* AND *LL* ARE IN MILLION LINKS.

	UL	**LL**	**AL**	**BFS**
$\mathcal{F} = 0.1$	161 M	165 M	1.02350	0.00411
$\mathcal{F} = 0.2$	1,099 M	1,152 M	1.04786	0.02868
$\mathcal{F} = 0.3$	3,778 M	4,198 M	1.11109	0.10452
$\mathcal{F} = 0.4$	8,347 M	10,246 M	1.22754	0.25508
$\mathcal{F} = 0.5$	15,130 M	21,558 M	1.42489	0.53669
$\mathcal{F} = 0.6$	21,687 M	37,003 M	1.70621	0.92117
$\mathcal{F} = 0.7$	29,224 M	64,521 M	2.20779	1.60623
$\mathcal{F} = 0.8$	35,294 M	106,211 M	3.00930	2.64408
$\mathcal{F} = 0.9$	38,685 M	157,160 M	4.06248	3.91241
$\mathcal{F} = 1.0$	39,914 M	204,354 M	5.11974	5.08729

2.1: Manual selection of features by part of speech tags

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
Bible	0.98	0.86	0.81	0.95	0.69	0.39	0.71	0.70	0.72	0.56	0.80	0.58
Middle Ages	0.98	0.88	0.93	0.95	0.79	0.42	0.81	0.71	0.79	0.49	0.84	0.52

2.2: Automatic selection of features by part of speech tags

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

	UL	LL	AL	BFS
n, a, v	26,732 M	45,260 M	1.69305	1.12673
n, a, v, t, m	27,606 M	51,091 M	1.85072	1.27189

3: Multi word features

Example sentence: A B C D E F

bigram shingling: (A B), (B C), (C D), (D E), (E F)

bigram hash-breaking: (A B), (C D), (E F)

Trigram shingling: (A B C), (B C D), (C D E), (D E F)

Trigram hash-breaking: (A B C), (D E F)

	UL	LL	AL	BFS
Tri	123 M	160 M	1.303	0.00399
Bi	2,531 M	3,030 M	1.197	0.07543
Word	35,294 M	10,6211 M	3.009	2.64408

3: Multi word features

Example sentence: A B C D E F

bigram shingling: (A B), (B C), (C D), (D E), (E F)

bigram hash-breaking: (A B), (C D), (E F)

Trigram shingling: (A B C), (B C D), (C D E), (D E F)

Trigram hash-breaking: (A B C), (D E F)

	UL	LL	AL	BFS
Tri	123 M	160 M	1.303	0.00399
Bi	2,531 M	3,030 M	1.197	0.07543
Word	35,294 M	10,6211 M	3.009	2.64408

Summary

- Complexity of text re-use can't be answered by parallelisation (except a squared increase of hardware is possible)

- Removing frequent words compresses the feature index but it needs to be removed too much in order to significantly boost the system while significantly decreasing the recall

- PoS tags help to compress the feature index while keeping acceptable results; however, tend to keep more frequent features in the analysis bringing no real performance benefit

- Multi word featuring brings necessary performance boosts while keeping results good

Further work: Which multi word features (subsets of words) are good multi word features?

Contacts

For more details:
http://www.gcdh.de/en/

Google group for Historical Text Re-use:
http://groups.google.com/group/historical-text-re-use

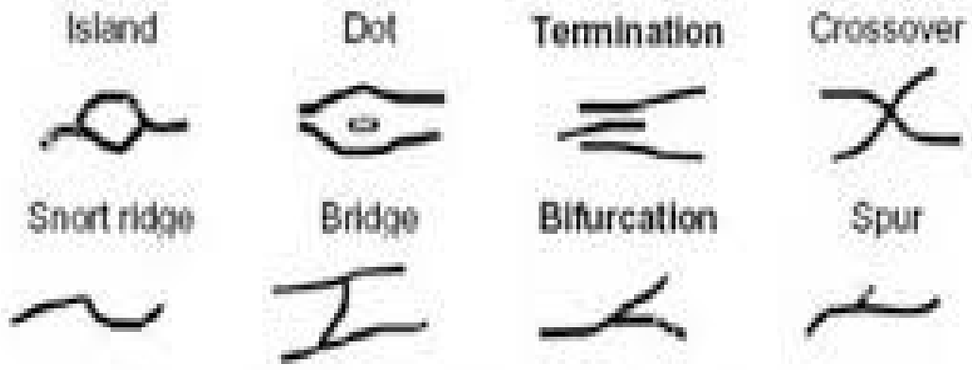
Marco Büchler
Göttingen Centre for Digital Humanities
Georg August University Göttingen, Germany
mbuechler@gcdh.de

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

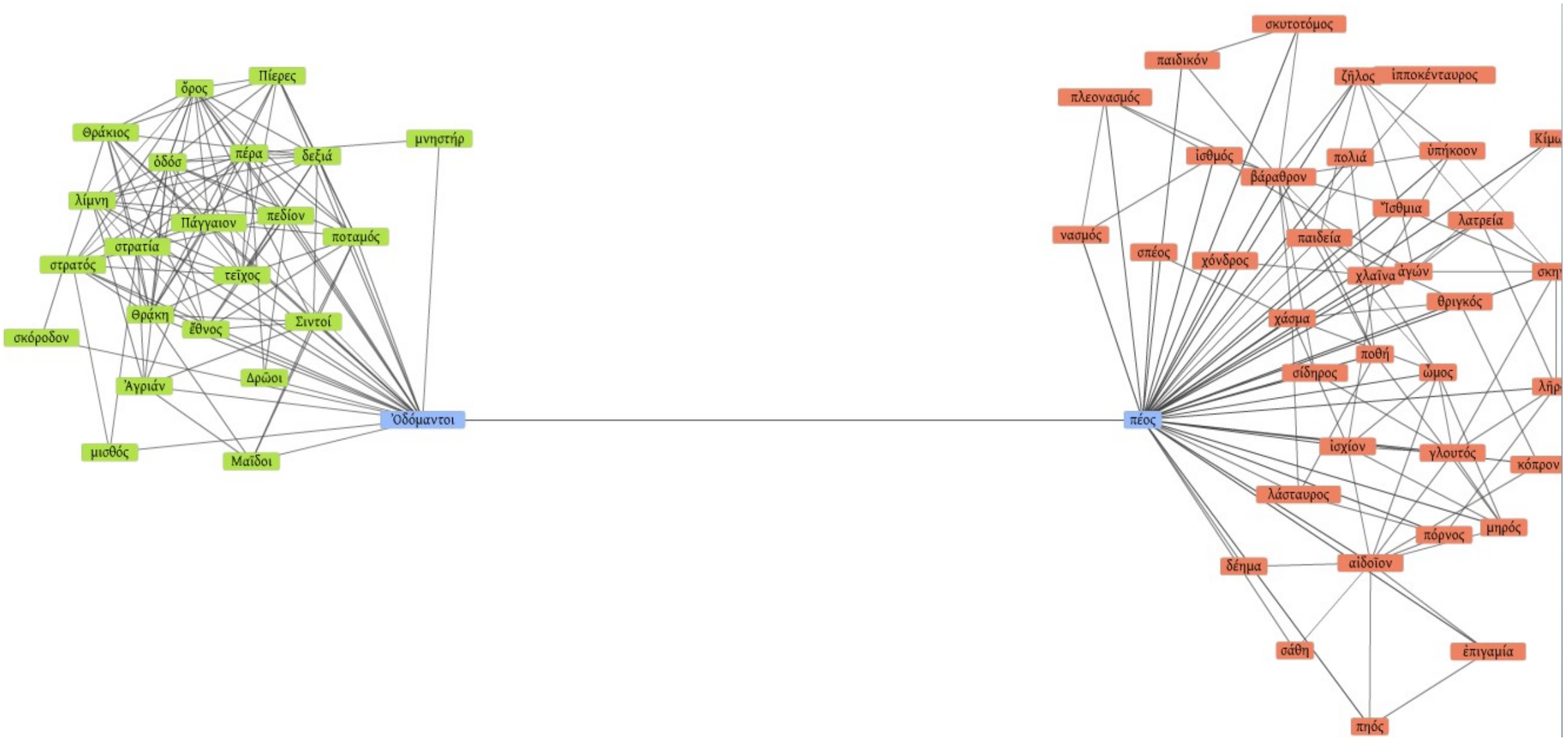
eTRAP:
Resulting questions

Question: What are the common primitives in the re-use diversity?

From biometry (Minutiae):



Identifying Passages of Interest in Text: Visualising Contrastive Semantics



Source: F. Baumgardt: Visualisierung von Kookkurrenzgraphen. Bachelorarbeit Abteilung Automatische Sprachverarbeitung, Universität Leipzig, 2010.

eTRAP (outline): Research focus

Minutiae: What are the common primitives of the re-use diversity?

Noisy Channel Mining: What is the system behind changes?

Understanding Re-use Diversity: What keeps stable and tends to be changed? (e. g. Influence of change of audience, sentiments etc)

Big Data view to textual criticism (Forensic Humanities): Profiling of author's re-use habits in order to ask questions like: Who tended to quote more literally than others?

Translation techniques (re-use style between languages): e. g. How does the translation style changes by different authors, languages, epochs etc.