ELECTRONIC TEXT REUSE ACQUISITION PROJECT

DETECTION OF TEXT REUSE IN HISTORICAL TEXTS

Marco Büchler, Greta Franzini, Maria Moritz, Emily Franzini, Gabriela Rotari
1. Definition & motivation

2. Research on the characteristics

3. Characteristics: Qualitative research

4. Characteristics: Quantitative research

5. Research on the reuse process


7. Process: Quantitative view

8. Results
DEFINITION & MOTIVATION
WHAT DO YOU ASSOCIATE WITH TEXT REUSE AND INTERTEXTUALITY?

this is true

this is true

this is truth

please consider this before talking/typing
**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.
WHO IS THIS PERSON?
“REUSE FROM SAME SOURCE”: COMMONALITIES & DIFFERENCES
**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“.
Evaluation of the reuse detection process by forensic criterions (standard in biometry):

- **Universality**: How universal 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 resistant 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
**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
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?**
Tasks: Verify presence of motifs in different collections and record their “base form” as text reuse training data.

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

Figure 2: Grimm motifs reduced to keywords.
Train an (adapted) **Named Entity Recognition (NER)** tagger, ideally as language-independent as possible, to **automatically annotate** further fairy tales and texts.
RQ: How to computationally detect a motif despite its variants?

For example:

- **DE [Grimm]**: *Schneewittchen und die sieben Zwerge*
- **EN [Briggs]**: *Snow White and the three robbers*
- **IT [Calvino]**: *Bella Venezia e i dodici ladroni*
- **SQ [von Hahn]**: *Schneewittchen und die vierzig Drachen*
- **RU [Pushkin]**: 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 (National Research Council Canada) Emotion Lexicon:

- The Roget Thesaurus
- 14,182 words types

**Emotions:** (Plutchik, 1980)
- anger
- anticipation
- disgust
- fear
- joy
- sadness
- surprise
- trust

**Sentiments:**
- negative emotions
- positive emotions
Classroom Questionnaires

- Empathy
- Identification
- Transportation

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

- Data set
CHARACTERISTICS: QUANTITATIVE RESEARCH
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.
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.
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.

RESULTS OF PARTICIPANTS

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

<table>
<thead>
<tr>
<th>Part of Speech-Tag</th>
<th>Wortartklasse</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>noun</td>
</tr>
<tr>
<td>v</td>
<td>verb</td>
</tr>
<tr>
<td>t</td>
<td>participle</td>
</tr>
<tr>
<td>a</td>
<td>adjective</td>
</tr>
<tr>
<td>d</td>
<td>adverb</td>
</tr>
<tr>
<td>l</td>
<td>article</td>
</tr>
<tr>
<td>g</td>
<td>particle</td>
</tr>
<tr>
<td>c</td>
<td>conjunction</td>
</tr>
<tr>
<td>r</td>
<td>preposition</td>
</tr>
<tr>
<td>p</td>
<td>pronoun</td>
</tr>
<tr>
<td>m</td>
<td>numeral</td>
</tr>
<tr>
<td>i</td>
<td>interjection</td>
</tr>
<tr>
<td>e</td>
<td>exclamation</td>
</tr>
<tr>
<td>u</td>
<td>punctuation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>v</th>
<th>t</th>
<th>a</th>
<th>d</th>
<th>l</th>
<th>g</th>
<th>c</th>
<th>r</th>
<th>p</th>
<th>m</th>
<th>u</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibel</td>
<td>0.98</td>
<td>0.86</td>
<td>0.81</td>
<td>0.95</td>
<td>0.69</td>
<td>0.39</td>
<td>0.71</td>
<td>0.70</td>
<td>0.72</td>
<td>0.56</td>
<td>0.80</td>
<td>0.58</td>
</tr>
<tr>
<td>Mittelalter</td>
<td>0.98</td>
<td>0.88</td>
<td>0.93</td>
<td>0.95</td>
<td>0.79</td>
<td>0.42</td>
<td>0.81</td>
<td>0.71</td>
<td>0.79</td>
<td>0.49</td>
<td>0.84</td>
<td>0.52</td>
</tr>
</tbody>
</table>
RESULTS OF PARTICIPANTS

- Inter-PoS analysis (dependencies between PoS tags)
- Interrater analysis
- Making data-sets available (including tagged data)
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.
• 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
### Table 1: Operation list for the automated approach

<table>
<thead>
<tr>
<th>operation</th>
<th>description</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOP(reuse_word, orig_word)</td>
<td>Original and reuse word are equal.</td>
<td>NOP(maledictus,maledictus)</td>
</tr>
<tr>
<td>upper(reuse_word, orig_word)</td>
<td>Word is lowercase in reuse and uppercase in original.</td>
<td>upper(kai,Kai) - in Greek</td>
</tr>
<tr>
<td>lower(reuse_word, orig_word)</td>
<td>Word is uppercase in reuse and lowercase in original.</td>
<td>lower(Gloriam,gloriam)</td>
</tr>
<tr>
<td>lem(reuse_word, orig_word)</td>
<td>Lemmatization leads to equality of reuse and original.</td>
<td>lem(penetrat,penetrabit)</td>
</tr>
<tr>
<td>repl_syn(reuse_word, orig_word)</td>
<td>Reuse word replaced with a synonym to match original word.</td>
<td>repl_syn(magnificavit,glorificavit)</td>
</tr>
<tr>
<td>repl_hyper(reuse_word, orig_word)</td>
<td>Word in Bible verse is a hyperonym of the reused word.</td>
<td>hyper(cupit,habens)</td>
</tr>
<tr>
<td>repl_hypo(reuse_word, orig_word)</td>
<td>Word in Bible verse is a hyponym of the reused word.</td>
<td>hypo(dederit,tollet)</td>
</tr>
<tr>
<td>repl_co-hypo(reuse_word, orig_word)</td>
<td>Reused word and original have the same hyperonym.</td>
<td>repl_co-hypo(magnificavit,fecit)</td>
</tr>
<tr>
<td>NOPmorph(reuse_tags, orig_tags)</td>
<td>Case or PoS did not change between reused and original word.</td>
<td>NOPmorph(na,na)</td>
</tr>
<tr>
<td>repl_pos(reuse_tag, orig_tag)</td>
<td>Reuse and original contain the same cognate, but PoS changed.</td>
<td>repl_pos(n,a)</td>
</tr>
<tr>
<td>repl_case(reuse_tag, orig_tag)</td>
<td>Reuse and original have the same cognate, but the case changed.</td>
<td>repl_case(g,d) - cases genitive, dative</td>
</tr>
<tr>
<td>lemma_missing(reuse_word, orig_word)</td>
<td>Lemma unknown for reuse or original word.</td>
<td>lemma_missing(tentari, inlectus)</td>
</tr>
<tr>
<td>no_rel_found(reuse_wword, orig_word)</td>
<td>Relation for reuse or original word not found in AGWN.</td>
<td>no_rel_found(gloria,arguitur)</td>
</tr>
</tbody>
</table>
PROCESS: QUANTITATIVE VIEW
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.
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].
chapter 1 it be a truth universally understand that a single man in possession of a good fortune must need a wife
so when a wealthy gentleman arrive in a neighbourhood it be clear that he must soon become the property of someone
daughter
my dear Mr. Bennet say he wife to he one day have you hear that someone have rent the house at Netherfield Park at last
Mr. Bennet reply that he have not
yes she continue Mrs. Long have just be here and she tell I all about it
Mr. Bennet do not answer
do you not want to know who have take it
cry he wife impatiently
you want to tell I and I have no objection to hear it
well Mrs. Long say that Netherfield have be take by a young man of large fortune from the north of England
what be he name
Bingley
be he marry or single
oh
single my dear
a single man of large fortune four or five thousand pound a year
what a fine thing for we girl

chapter 1 it be a truth universally acknowledge that a single man in possession of a good fortune must be in want of a wife
however little known the feeling or view of such a man may be on he first enter a neighbourhood this truth be so well fix in the mind of the surround family that he be consider the rightful property of some one or other of they daughter
my dear Mr. Bennet say he lady to he one day have you hear that Netherfield Park be let at last
Mr. Bennet reply that he have not
but it be return she for Mrs. Long have just be here and she tell I all about it
Mr. Bennet make no answer
do you not want to know who have take it
cry he wife impatiently
you want to tell I and I have no objection to hear it
this be invitation enough
why my dear you must know Mrs. Long say that Netherfield be take by a young man of large fortune from the north of England that he come down on Monday in a chaise and tour to see the place and be so much delight with it that he agree with Mr. Morris immediately that he be to take possession before Michaelmas and some of the servant be to be in the house by the end of next week

what be he name
Bingley
be he marry or single
oh
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:

\[ R_{ARI} = 4.71 \left( \frac{\text{characters}}{\text{words}} \right) + 0.5 \left( \frac{\text{words}}{\text{sentences}} \right) - 21.43 \]  

(1)

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

\[ R_{DCI} = 0.1579 \left( \frac{\text{difficult words}}{\text{words}} \times 100 \right) + 0.0496 \left( \frac{\text{words}}{\text{sentences}} \right) \]

(2)
RESULTS OF THE SIMPLIFICATION WITH READABILITY TESTS

Readability test result matrix:

<table>
<thead>
<tr>
<th></th>
<th>ARI</th>
<th>Dale-Chall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Novel</td>
<td>14-15 year olds</td>
<td>14-16 year olds</td>
</tr>
<tr>
<td>Graded Reader</td>
<td>11-12 year olds</td>
<td>11-13 year olds</td>
</tr>
</tbody>
</table>
An example of a structural text simplification $\rightarrow$ many-to-one.
COMPARISON OF WORD LENGTH

Word length distribution

Original text
Graded reader

Length of word
Probability
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

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>upon</td>
<td>75</td>
<td>table</td>
<td>31</td>
</tr>
<tr>
<td>least</td>
<td>65</td>
<td>astonishment</td>
<td>30</td>
</tr>
<tr>
<td>acquaintance</td>
<td>63</td>
<td>fancy</td>
<td>30</td>
</tr>
<tr>
<td>either</td>
<td>59</td>
<td>attempt</td>
<td>29</td>
</tr>
<tr>
<td>whose</td>
<td>59</td>
<td>dine</td>
<td>29</td>
</tr>
<tr>
<td>dare</td>
<td>53</td>
<td>beg</td>
<td>28</td>
</tr>
<tr>
<td>regard</td>
<td>53</td>
<td>depend</td>
<td>28</td>
</tr>
<tr>
<td>determine</td>
<td>47</td>
<td>highly</td>
<td>28</td>
</tr>
<tr>
<td>scarcely</td>
<td>45</td>
<td>satisfaction</td>
<td>28</td>
</tr>
<tr>
<td>ladyship</td>
<td>42</td>
<td>acknowledge</td>
<td>27</td>
</tr>
<tr>
<td>former</td>
<td>38</td>
<td>credit</td>
<td>27</td>
</tr>
<tr>
<td>put</td>
<td>36</td>
<td>thus</td>
<td>27</td>
</tr>
<tr>
<td>amiable</td>
<td>35</td>
<td>disposition</td>
<td>26</td>
</tr>
<tr>
<td>deal</td>
<td>34</td>
<td>exceedingly</td>
<td>26</td>
</tr>
<tr>
<td>design</td>
<td>32</td>
<td>praise</td>
<td>26</td>
</tr>
</tbody>
</table>
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
“Salvation for the Rich”
Clement of Alexandria
Christian theologian, 2nd cent.

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

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

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 Bibliindex team (Mellerin, 2014; Mellerin, 2016)
- We obtain 162 verse-reuse-pairs
- Pointing to 31 Bible books
## BIBLICAL REUSE EXAMPLES

<table>
<thead>
<tr>
<th>more literal</th>
<th>Bible verse</th>
<th>Bernard reuse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proverbs 18:3</td>
<td>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)</td>
<td>Impius , cum venerit in profundum malorum , contemnit (When the wicked man is come into the depth of evil)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>less literal</th>
<th>Bible verse</th>
<th>Clement reuse</th>
</tr>
</thead>
</table>
| 1Cor 13:13   | νυνὶ δὲ μένει πίστις , ἐλπίς , ἁγάπη , τὰ τρία ταῦτα μεῖζων δὲ τούτων ἢ ἁγάπη (And now remain faith, hope, love, these three; but the greatest of those is love.) | πίστει καὶ ἐλπίδι καὶ ἁγάπῃ (faith, and hope, and love - in dative case)  
ἀγάπην , πίστιν , ἐλπίδα (love, faith, hope - in accusative case)  
μένει δὲ τὰ τρία ταῦτα , πίστις , ἐλπίς , ἁγάπη · μεῖζων δὲ ἐν τούτοις ἢ ἁγάπη (and remain these three, faith, hope, love; but the greatest among them is love) |

<table>
<thead>
<tr>
<th>non-literal</th>
<th>Bible verse</th>
<th>Clement reuse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mt 12:35</td>
<td>ὁ ἁγαθὸς ἄνθρωπος ἐκ τοῦ ἁγαθοῦ θησαυροῦ ἐκβάλλει ἁγαθά , 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 .)</td>
<td>Ψυχῆς , τὰ δὲ ἐκτὸς , κἂν μὲν ἡ ψυχὴ χρῆται καλῶς , καλὰ καὶ ταῦτα δοκεῖ , ἐὰν δὲ πονηρῶς , πονηρᾶ , 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)</td>
</tr>
</tbody>
</table>
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/]
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.
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.

<table>
<thead>
<tr>
<th></th>
<th>literal</th>
<th></th>
<th>non-literal</th>
<th></th>
<th>unclassified</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NOP</td>
<td>upper</td>
<td>lower</td>
<td>lem</td>
<td>syn</td>
<td>hyper</td>
</tr>
<tr>
<td>Greek</td>
<td>337</td>
<td>6</td>
<td>0</td>
<td>356</td>
<td>153</td>
<td>20</td>
</tr>
<tr>
<td>Latin</td>
<td>587</td>
<td>0</td>
<td>44</td>
<td>102</td>
<td>60</td>
<td>14</td>
</tr>
</tbody>
</table>
Operations that successfully looked up a lemma:

\[ \text{lem_success} = \{\text{lem, syn, repl_hyper, repl_hypo, repl_co-hypo, no_rel_found}\} \]

with \text{lem_missing} representing not found tokens in the lemmata.

\[
\text{cov}_{\text{lem}} = \frac{\sum_{\text{o} \in \text{lem_success}} \text{Occ}(\text{o})}{\sum_{\text{o} \in \text{lem_success} \cup \{\text{lem_missing}\}} \text{Occ}(\text{o})}
\]

\[
\text{cov}_{\text{AGWN}} = \frac{\sum_{\text{o} \in \text{agwn_success}} \text{Occ}(\text{o})}{\sum_{\text{o} \in \text{agwn_success} \cup \{\text{no_rel_found}\}} \text{Occ}(\text{o})}
\]

We obtain a \text{cov}_{\text{lem}} of \textbf{0.65} for our Greek and \textbf{0.88} for the Latin data-set. And a \text{cov}_{\text{AGWN}} of \textbf{0.34} for our Greek and \textbf{0.33} for our Latin data-set.

Language resources help to get an idea of reuse components.
Visit us

🌐 http://www.etrap.eu
✉️ contact@etrap.eu

*Stealing from one is plagiarism, stealing from many is research*

*(Wilson Mitzner, 1876-1933)*
The theme this presentation is based on is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. Changes to the theme are the work of eTRAP.