

BROTHERS GRIMM, JANE AUSTEN AND PAULUS OROSIUS HAVE ONE THING IN COMMON: THE eTRAP RESEARCH TEAM AND ITS DH PROJECTS

Emily Franzini and Greta Franzini

February 23, 2017



1. Who are we?
2. What is text reuse?
3. Our research

INTERRUPT US AT ANY TIME!

WHO ARE WE?

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.

WHAT IS TEXT REUSE?

Text reuse = spoken and written repetition of text across time and space.

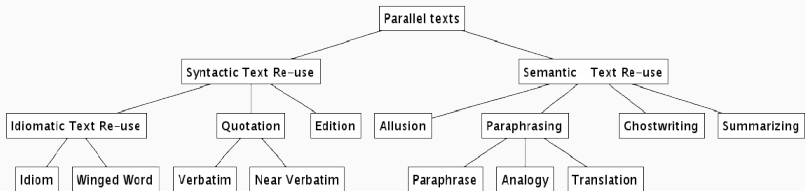


Figure 1: Text reuse styles.

Question:

Why is text reuse detection relevant for Humanities and Computer Science?

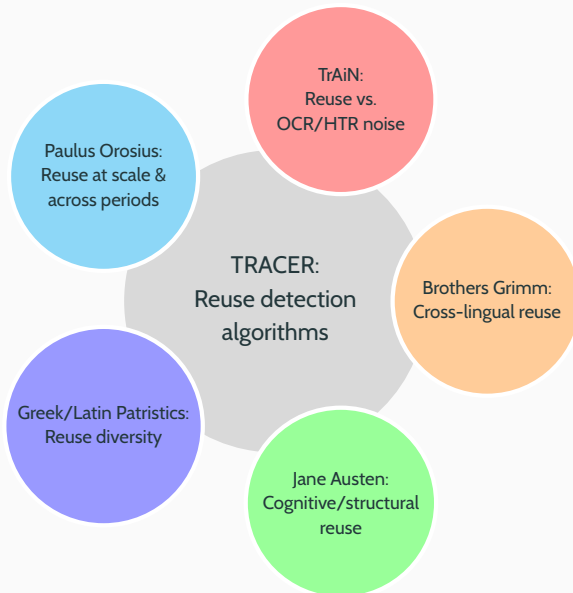
- **Humanities:**
 - Lines of transmission and textual criticism.
 - Transmissions of ideas & thoughts under different circumstances and conditions.
- **Computer Science:**
 - Text decontamination for stylometry and authorship attribution, dating of texts.
 - Text Mining, Corpus Linguistics.

Text reuse challenges:

- Detecting text reuse **at scale** (Big Data: information overload vs. information poverty);
- Detecting text reuse across **languages**;
- Detecting **looser forms** of text reuse, e.g. allusion;
- **Diversity** of **historical** texts: language evolution, copy errors, etc.

OUR RESEARCH

OVERVIEW OF OUR PROJECTS: HISTORICAL DATA





TRACER

TEXT REUSE DETECTION MACHINE

TRACER: OVERVIEW

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

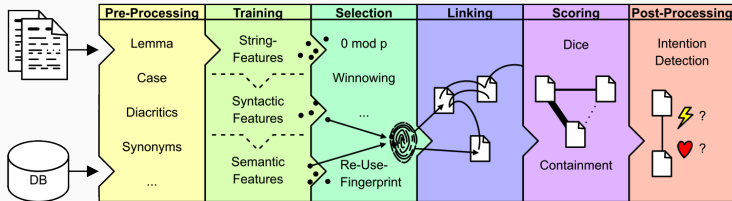


Figure 2: Detection task in six steps. More than 1M permutations of implementations of different levels are possible.

TRACER is language-independent.

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

TEXT REUSE INVESTIGATION INTO THE FIRST CHRISTIAN HISTORY OF ROME

PAULUS OROSIUS AND HIS HISTORIES

Paulus Orosius [ca. AD 375-418]

- Roman historian and a Christian from Spain;
- Student of St Augustine [AD 354-430].

Historiae adversus Paganos = *Histories against the Pagans**

- First Christian history of Rome;
- Complementary to St Augustine's *De civitate Dei contra Paganos*;
- Defense against pagan accusations that Rome's decline was caused by the advent of Christianity;
- Heavily reuses both pagan and Christian authors to reject pagan claims.

*Paganism = pantheism, polytheism, non-Christian.

*Christianity = monotheism. Declared *permitted religion* by Constantine the Great in 313 (Edict of Milan); declared official religion of the Empire by son Constantius II in 350.

- **Close Reading**

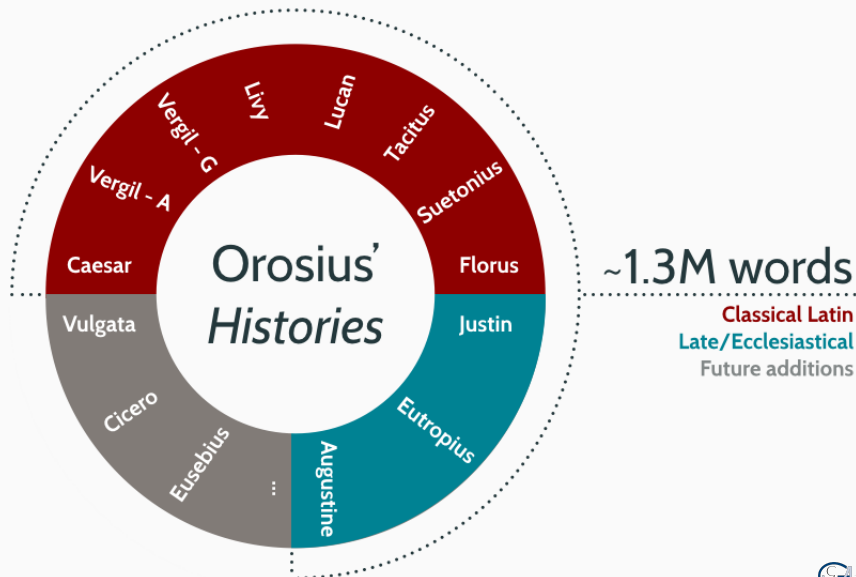
How does Orosius **reuse text** in order to build his defense?

- **Close + Distant Reading**

Can we quantify and categorise Orosius' **reuse diversity** (taxonomy)?

- **Distant Reading**

How does a **large, diachronic corpus** affect **automatic text reuse detection** and its **performance**?



Orosius:

- reuses **two words to entire sentences or even paragraphs**;
- quotes **word-for-word** (i.e. *verbatim*), near-verbatim **or (very) loosely**;
- doesn't always cite the original author;
- occasionally **misattributes words** because citing from memory;
- reuses **text that doesn't survive**.

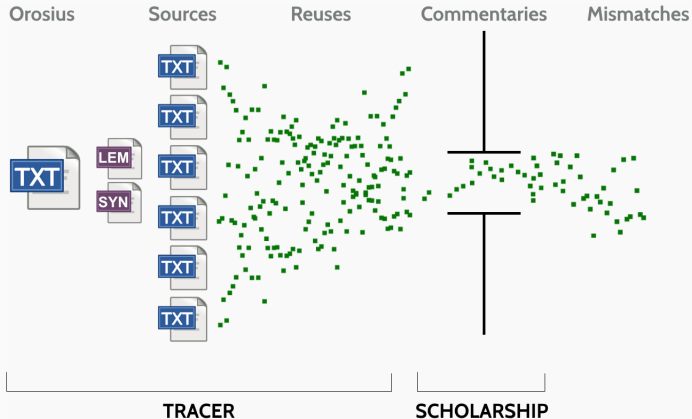
*Nec tibi cura canum fuerit **postrema*** (Georg. 3.404 - poetry)

[= Nor be your dogs last cared for]

*non est tamen **canum cura postrema*** (Oros. 1.1. - prose)

[= Dogs are not to be cared for last]

METHODOLOGY: TRACER VS. COMMENTARIES



- How do the **computed results** compare to **existing scholarship**?
- Has TRACER identified reuses that existing scholarship hasn't?
- Has existing scholarship identified reuses that TRACER hasn't?

RESULTS: OROSIUS' REUSE OF TACITUS' HISTORIAE

- Reuses documented in **primary sources** (*precision*): **15** (= 12+3?)
- Reuses TRACER can attempt to match: **5** (10 are of lost text)
- Reuses identified by **TRACER** (*recall*)*: **40**
 - **verbatim**;
 - **near-verbatim**: “true” and “false” (spelling conventions);
 - **no similarity**: *why?* Synonym replacement? PoS? Feature density?

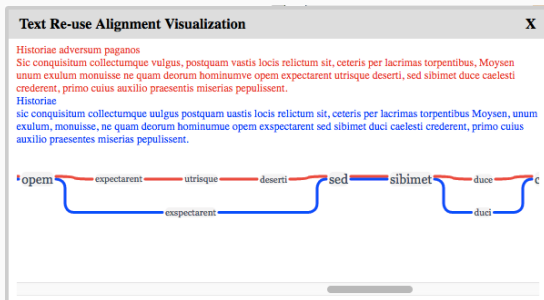
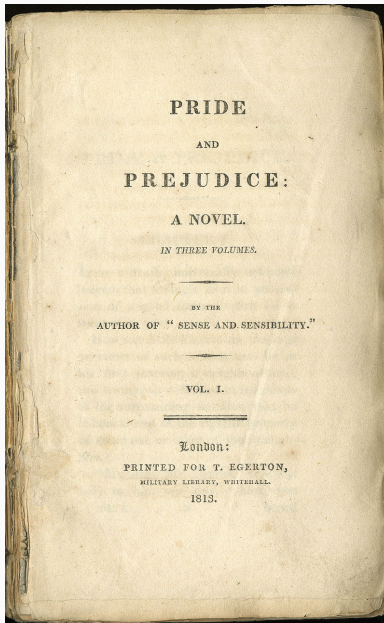
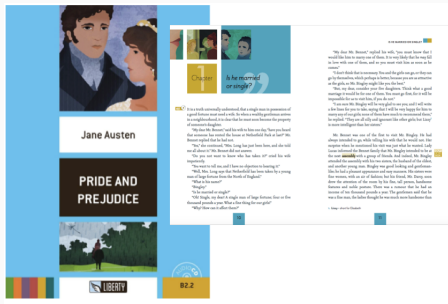


Figure 3: Orosius 1.10, Tacitus 5.3.

JANE AUSTEN AND TEXT SIMPLIFICATION

JANE AUSTEN'S PRIDE & PREJUDICE



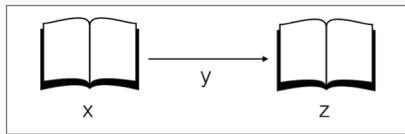


Definition:

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].

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

Structural changes:

- Elizabeth is **exceedingly handsome**.
- Elizabeth is **very beautiful**.

Cognitive changes:

- ... **Soon after this event**, Elizabeth received a visit...

Structural & cognitive changes:

- Elizabeth is **exceedingly beautiful**.

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**:

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

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

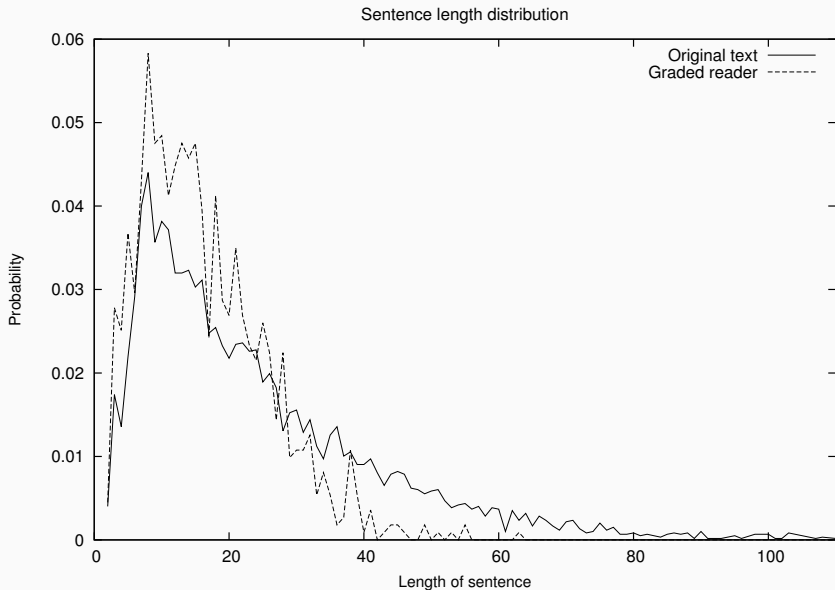
$$R_{DCI} = 0.1579 \left(\frac{\text{difficult words}}{\text{words}} * 100 \right) + 0.0496 \left(\frac{\text{words}}{\text{sentences}} \right) \quad (2)$$

RESULTS OF THE SIMPLIFICATION WITH READABILITY TESTS

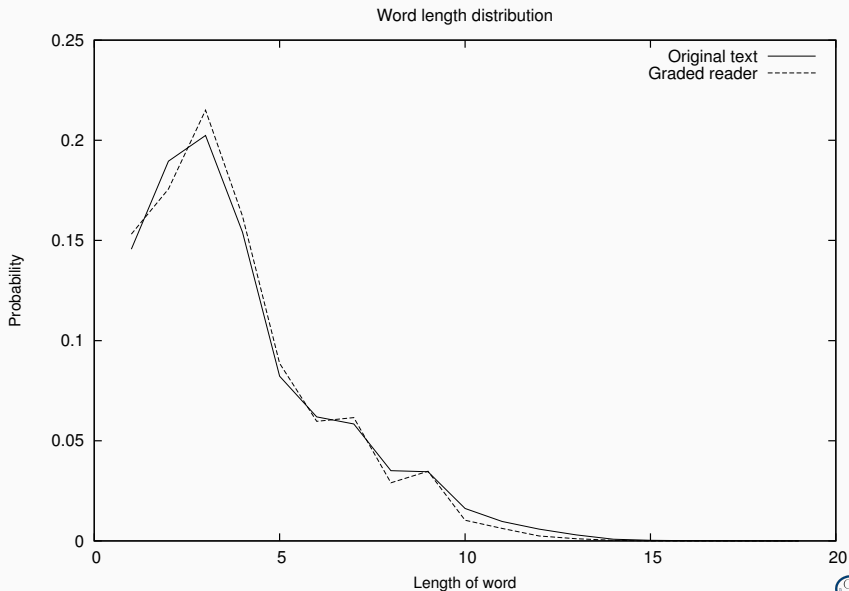
Readability test result matrix:

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

COMPARISON OF SENTENCE LENGTH



COMPARISON OF WORD LENGTH



SIMPLIFICATION & SENTENCE LENGTH

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

Text Re-use Alignment Visualization

X

GR

chapter 1 it be a truth universally understand that a single man in possession of a good fortune must need a wife

ON

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



AUTOMATIC ALIGNMENT OF ORIGINAL NOVEL WITH GRADED READER

378 Text Re-uses



GR

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

120000001

120000002

120000003

120000004

120000005

120000006

120000007

120000008

120000009

120000010

120000011

120000012

120000013

120000014

120000015

120000016

120000017

ON

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 four to see the place and be so much delighted with it that he agree with Mr. Morris immediately that he be to take possession before Michaelmas and some of he servant be to be in the house by the end of next week

what be he name

Bingley

be he marry or single

oh

130000001

130000002

130000003

130000004

130000005

130000006

130000007

130000008

130000009

130000010

130000011

130000012

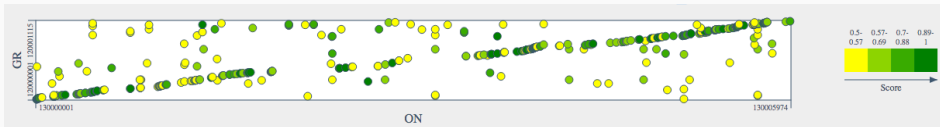
130000013

130000014

130000015

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:



TRACING AUTHORSHIP IN NOISE

“It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data (Dasu and Johnson, 2003). Data preparation is not just a first step, but must be repeated many times over the course of analysis as new problems come to light or new data is collected.” (Wickham, 2014, p. 1)

Duration: 6 months

Budget: 20,000€

Funder: University of Göttingen, Campuslabor Digitalisierung

Final expert workshop (March 2017): text reuse meets stylometry

Research Question:

When does Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) noise begin to interfere with automatic text reuse and authorship detection?

Case study: Correspondence of Brothers Grimm

CONCEPT: GRIMM CORRESPONDENCE



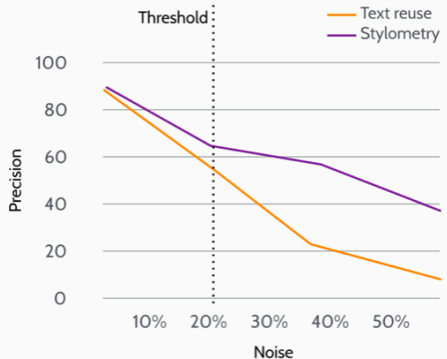
Digitised letters written by Jacob & Wilhelm Grimm from childhood to death.



H. Rölleke's 2001 print edition of the letters.

HTR Version
Transkribus

OCR Version
Ocropy & Tesseract





DIGITAL BREADCRUMBS *of*
BROTHERS GRIMM

The collection and automatic detection of folktale motifs as text reuse units across languages and traditions.

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: 185098476	Calvino_1956 VIAF: 181208131	Jacobs_1892 VIAF: 315397813	Bruford_1994 VIAF: 12471835	Hoogasian- Villa_1966 VIAF: 186329063	Campbell_1958 VIAF: 25969242	Taylor_1823 VIAF: 59071527	Briggs_1970 VIAF: 46803237	El-Shamy_1969 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 4: 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 5: Grimm motifs reduced to keywords.

Thompson Motif Index (TMI) ontology (OWL/RDF), by Antónia Koštová, Thierry Declerck and Tyler Klement (Declerck et al., 2016).

```
<http://www.semanticweb.org/tonka/ontologies/2015/5/tmi-atu-ontology#T11.6>
rdf:type :Motif ;
rdf:type :T11 ;
rdf:type owl:NamedIndividual ;
rdfs:comment "\"Terminal motif T11.6\""@en ;
rdfs:label "\"Wish for wife red as blood, white as snow, black as raven.\""@en ;
.
```

Figure 6: Representation of a motif in the TMI ontology. Image reproduced with permission of Thierry Declerck.

Visit us



<http://www.etrp.eu>



contact@etrp.eu

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



GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN



SPONSORED BY THE

Federal Ministry
of Education
and Research

- 1. Grimm (1812-1857) *Kinder- und Hausmärchen*.
- 2. Briggs, K. M. (1970) *A Dictionary of British Folk-Tales in the English Language: Part A: Folk Narratives*. London: Routledge & Kegan Paul.
- 3. Calvino, I. (1956) *Fiabe Italiane*. Mondadori.
- 4. Hahn, J. G. von (1864) *Griechische und Albanesische Märchen*, Zweiter Theil. Leipzig: Engelmann, pp. 137.
- 5. Пушкин, Александр Сергеевич (1799-1837). Сказка о мертвой царевне и о семи богатырях. Available at:
<http://rvb.ru/pushkin/01text/03fables/01fables/0800.htm> (Accessed: 27 June 2016).

REFERENCES

- Büchler, M. *TRACER: Text Reuse Detection Machine*. At: <http://www.etrapp.eu/research/tracer/>
- Dasu T., Johnson T. (2003) *Exploratory Data Mining and Data Cleaning*. John Wiley & Sons.
- Waring, R. (2012) *Writing graded readers*. At: <http://www.er-central.com/authors/writing-a-graded-reader/writing-graded-readers-rob-waring/>
- Wickham, H. (2014) 'Tidy Data', *Journal of Statistical Software*, 59(10). At: <https://www.jstatsoft.org/article/view/v059i10/v59i10.pdf>

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.

