DIGITAL BREADCRUMBS of BROTHERS GRIMM

Emily Franzini & Greta Franzini

July 15, 2016

Digital Humanities 2016, Kraków

TABLE OF CONTENTS

INTRODUCTION

Electronic Text Reuse Acquisition Project (eTRAP)

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

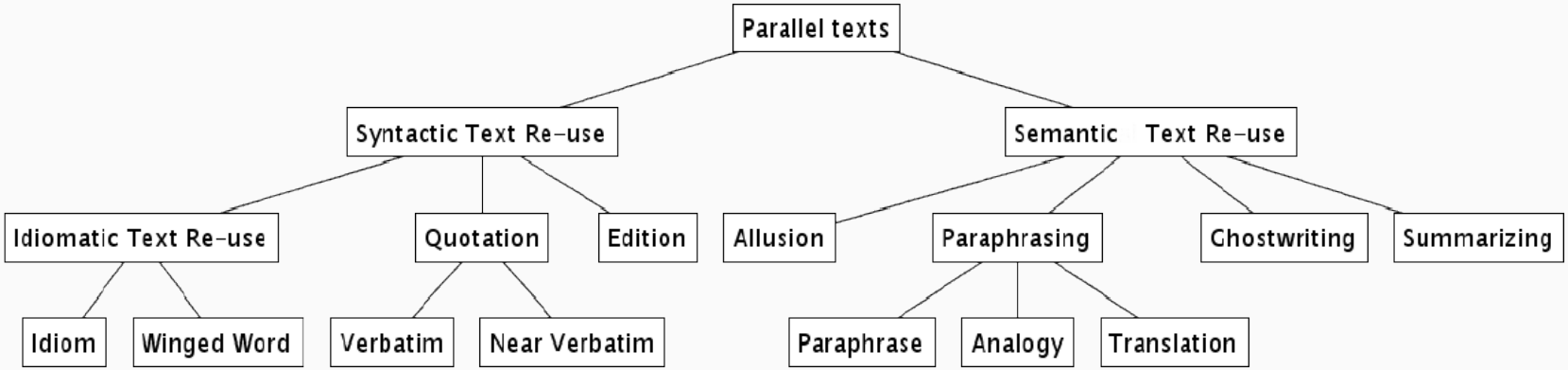

Figure 1: Text reuse styles [Author: Marco Büchler].

Specific interests: text reuse detection at scale (Big Data) and historical text reuse.

TRACER: suite of 700 algorithms; developed by Marco Büchler.

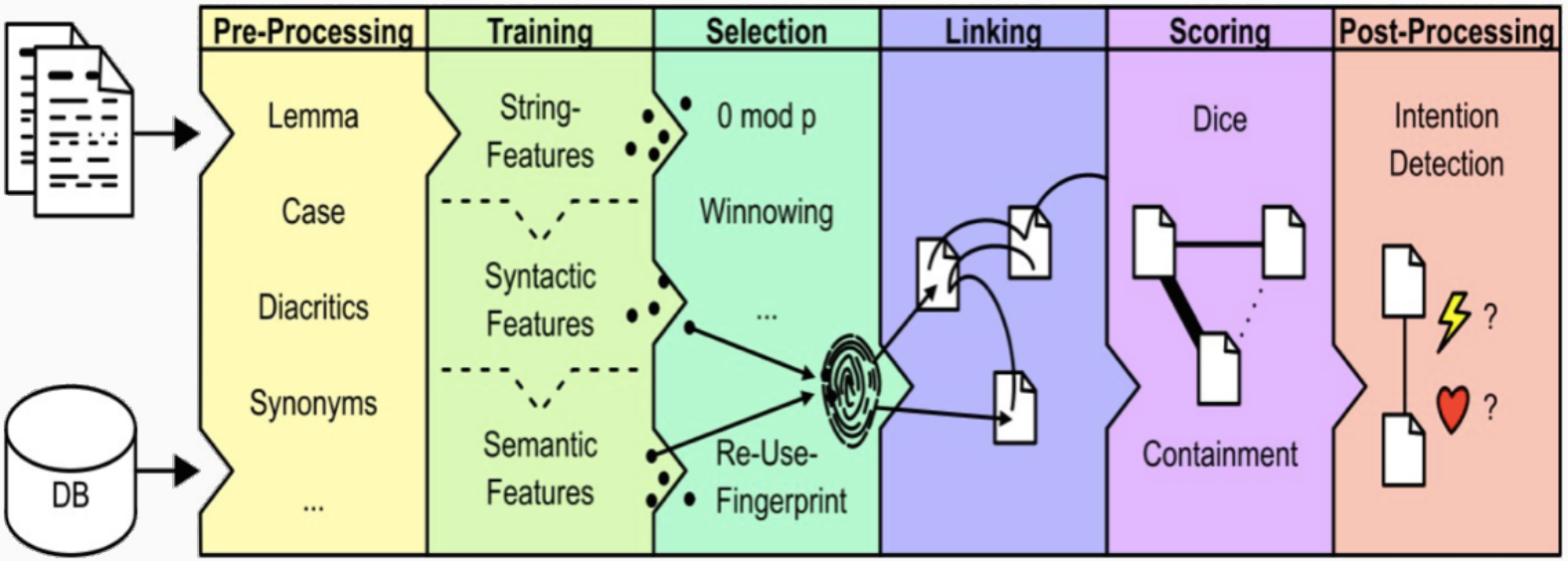

Figure 2: TRACER steps. More than 1M permutations of implementations of different levels are possible.

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

RESEARCH FOCUS

CURRENT CHALLENGES

Text reuse challenges:

- Detecting text reuse across languages;
- Detecting text reuse at scale;
- Detecting looser forms of text reuse, e.g. allusion;
- Diversity of historical texts: language evolution, copy errors, etc.

RESEARCH CONTRIBUTION

Our contribution:

- Advance research into cross-lingual detection;
- Testing stability and performance of text reuse detection at scale.

What this research is about:

- **Supporting** intertextual and folkloristic studies;

What this research is not about:

- **Studying** intertextual transmission.

COMPUTATIONAL FOLKLORISTICS

> *"Over the course of the past decade [...] the size and scope of digital archives of folklore have exploded, and the magnitude of digital materials available for folkloristic consideration has increased exponentially." (Tangherlini, 2016, p. 5).*

> *"We are in the **very early days of working computationally** with rich folklore resources [...]." (Tangherlini, 2016, p. 10).*

Tangherlini (2013) outlines four areas of research in computational folkloristics: (1) collecting **and archiving**, (2) **indexing and classifying**, (3) **visualization and navigation**, and (4) analysis.

RESEARCH DATA-SET

DIGITAL BREADCRUMBS OF BROTHERS GRIMM

Project began in October 2015.

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

Changes in:

- Size: from 156 to 211.
- Content: gruesome to mild.
- Style: Jacob scholarly, Wilhelm figurative.
- Language: Variants, diachronic evolution.

MOTIFS AS REUSE UNITS

Motif: *"1. A minimal thematic unit"* (Prince, 2003, p. 55), a measurable primitive.

Measurable primitives from an interdisciplinary standpoint:

- **Literature**: tracing MOTIFS
- **Cultural Studies**: tracing MEMES
- **Linguistics**: tracing PATTERNS
- **Computer Science**: tracing FEATURES
- **Forensics**: tracing MINUTIAE

GOAL

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

Tales selected for investigation:

- *Snow White* (AT 709);
- *Puss in Boots* (AT 545B);
- *The Fisherman and his Wife* (AT 555).

EXAMPLE CASE STUDY: SNOW WHITE

Q: 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 need to combine Aarne-Thompson (Uther) and Propp approaches. That is, finding the balance between describing a motif (AT specificity) and leaving enough space for variations (Propp typological unity and sequence of events).

Collections and Languages

- **Identified versions**: Albanian, Algerian, Appalachian, Armenian, Breton, Celtic (Scottish), Egyptian, English, Finnish, German, Greek, Italian, Moroccan, Russian, Spanish.
- **Potential others**: African, Australian, Basque, Caribbean, Catalan, Caucasian, Chinese, Danish, Dutch, Estonian, French, Friesian, Georgian, Hawaiian, Icelandic, Indian, Indian-American, Israeli, Japanese, Korean, Latvian, Lithuanian, Macedonian, Mexican, Nepalese, New Zealand, Norwegian, Paraguayan, Persian, Polish, Portuguese, Punjabi, Romansh, Rumanian, Siberian, South-American, Sri Lankan, Swedish, Swiss, Tibetan, Turkish, Uzbek, Yiddish.
- **Does not appear in**: Ladin.

DATA COLLECTION AND CURATION

Tasks: Verify presence of motif in different collections and record its ”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 3: 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 4: Grimm motifs reduced to keywords.

Premise: To trace a motif through space and time you need big data.

Table 1: Google Custom Search vs. Apache Lucene.

Approach	PROs	CONs
Google Custom Search (online)	-Huge data -API	-Not free -Limited result-set (top 100)
Apache Lucene (offline)	-Free -Control over search parameters	-Download & index all docs

Current research on online vs. offline approaches for text reuse detection (German idioms) at scale (Solhdoust, 2016):

- **Google Custom Search** (online): searching in Google Books and the web.
- **Apache Lucene** (offline): searching in Deutsches Textarchiv, zeno.org, Project Gutenberg.

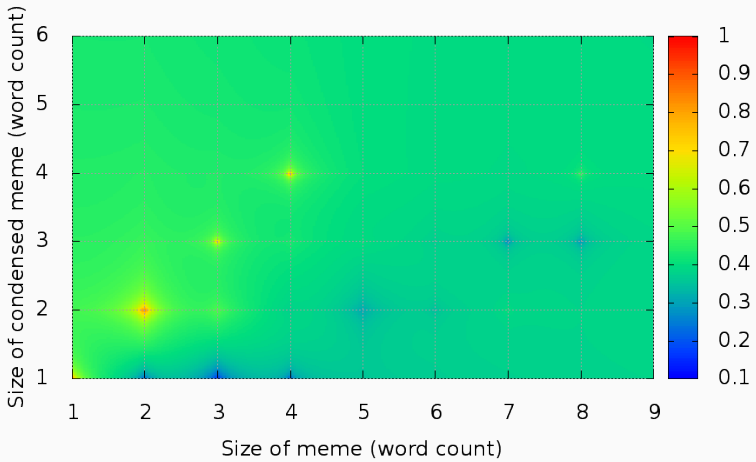

Figure 5: Similarity plot of idiom/meme samples using Google's Custom Search engine (online).

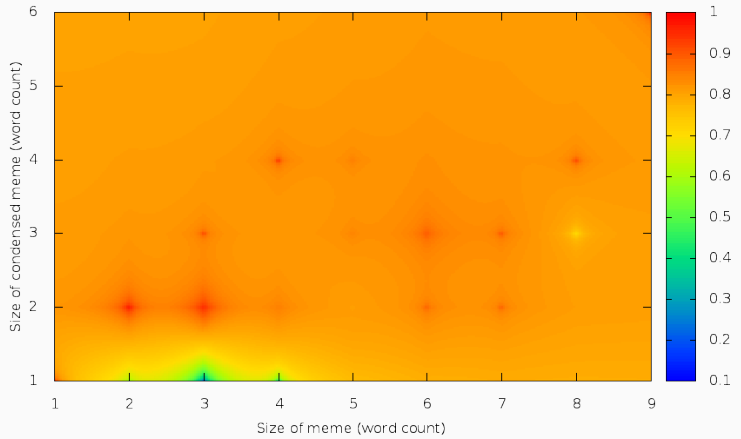

Figure 6: Similarity plot of idiom/meme samples using Apache Lucene (offline).

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

CONCLUSION

CONCLUSION

Contribution so far:

- Multilingual, curated dataset (not openly available yet);
- Results for online vs. offline text reuse detection at scale.

Short-term objectives:

- Run computational analyses on collected folktale data and study the results;
- Release multilingual dataset in SKOS XL for integration with existing ontological resources;
- Extend dataset to more languages and collections.

Thank you.
Questions?

Presentation

Greta Franzini and Emily Franzini

Team (in alphabetical order)

Marco Büchler, Emily Franzini, Greta Franzini, Franzi Pannach, Gabriela Rotari.

Visit us


http://www.etrap.eu


etrap@gcdh.de

eTRAP
Electronic Text Reuse Acquisition Project
INSTITUTE OF COMPUTER SCIENCE
GÖTTINGEN CENTRE FOR DIGITAL HUMANITIES

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

SPONSORED BY THE

Federal Ministry
of Education
and Research

NOTES

- 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

- Declerck, T., Koštová, A., Klement, T. (2016) *Ontologisierung vom Thompson Motif's Index*, Poster. Digital Humanities im deutschsprachigen Raum Konferenz, 7-12 March, Leipzig.
- Prince, G. (2003) *A Dictionary of Narratology*. University of Nebraska Press.
- Solhdoust, M. (2016) *Text Reuse at Web-scale*. Master's Thesis. University of Göttingen.
- Tangherlini, T. R. (2016) 'Big Folklore: A Special Issue on Computational Folkloristics', *Journal of American Folklore*, 129(511), pp. 5-13 [Online]. DOI: 10.1353/jaf.2016.0000
- Tangherlini, T. R. (2013) 'The Folklore Macroscope: Challenges for a Computational Folkloristics. The 34th Archer Taylor Memorial Lecture', *Western Folklore*, 72(1), pp. 7-27 [Online]. Available at: `http://tango.bol.ucla.edu/publications/A99.pdf` (Accessed: 25 June 2016).

APPENDIX

IMAGES USED

Public domain

- Grimm profile.
  At: `https://commons.wikimedia.org/wiki/File%3ABr%C3%BCder_Grimm_Doppelportr%C3%A4t_1843.jpg` (Accessed: 25 June 2016).

Proprietary

- Google & Lucene plots by Mahdi Solhdoust.

LICENCE

The LaTeX 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.