Digital Breadcrumbs of Brothers Grimm

Emily Franzini & Greta Franzini
July 15, 2016
Digital Humanities 2016, Kraków
1. Introduction

2. Research Focus

3. Research Data-set

4. Conclusion

5. Appendix
INTRODUCTION
**Electronic Text Reuse Acquisition Project (eTRAP)**

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

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
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.
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.
"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
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.
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
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).
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.
- **Does not appear in**: Ladin.
Tasks: Verify presence of motif in different collections and record its "base form" as text reuse training data.

Figure 3: Microsoft Excel matrix of motifs. Left column lists AT motifs in *Snow White* (AT 709); top row lists languages and collections covered.

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.

<table>
<thead>
<tr>
<th>Approach</th>
<th>PROs</th>
<th>CONs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Google Custom Search</strong> (online)</td>
<td>- Huge data</td>
<td>- Not free</td>
</tr>
<tr>
<td></td>
<td>- API</td>
<td>- Limited result-set (top 100)</td>
</tr>
<tr>
<td><strong>Apache Lucene</strong>       (offline)</td>
<td>- Free</td>
<td>- Download &amp; index all docs</td>
</tr>
<tr>
<td></td>
<td>- Control over search parameters</td>
<td></td>
</tr>
</tbody>
</table>
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).
Integration with existing resources

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

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
• 1. Grimm (1812-1857) *Kinder- und Hausmärchen*.


APPENDIX
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.
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.