GRIMMS MÄRCHEN

AN INTERTEXTUAL PUZZLE OF FAIRY TALES

Greta Franzini

February 23, 2016

National University of Ireland, Galway

eTRAP
Electronic Text Reuse Acquisition Project
GÖTTINGEN CENTRE FOR
DIGITAL HUMANITIES

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

TABLE OF CONTENTS

INTRODUCTION

Electronic **T**ext **R**euse **A**cquisition **P**roject (eTRAP)

Early Career Research Group funded by German Ministry of Education & Research (BMBF).

Budget: €1.7M.
Duration: January 2015 – February 2019. Research since October 2015.
Team: 4 core staff; 9 student assistants.

- **Interdisciplinary**: Classics, Computer Science, German Studies, Mathematics, Philosophy, Software Engineering.
- **International**: 7 nationalities, 11 languages spoken.

Professional team coaching for effective group dynamic:

- Effective communication;
- Making the most of strengths;
- Effective delegation.

coach evolve lead
inspeyer

RESEARCH FOCUS

Electronic Text Reuse Acquisition Project (eTRAP)

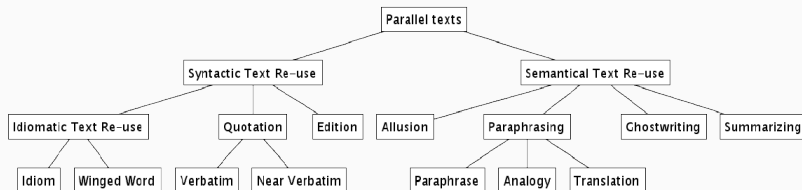Text reuse = spoken and written repetition of text across time and space.

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

Current limitations:

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

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

DIGITAL HUMANITIES

Ulrike Rieß (*Big Data bestimmt die IT-Welt*):

- Large amounts of data that can't be processed and analysed manually;
- Less structured data, e.g. in comparison to databases and data warehouse systems;
- Linked data between heterogeneous and distributed resources.

Information overload = large amounts of data (Big Data).
Information poverty = noisy, missing, fragmentary, oral data (Humanities Data).

COMPLEXITY

Inconsistent survival, documentation, cataloguing.

COMPLEXITY

HISTORICAL TEXT REUSE

Historical text reuse as an opportunity to tackle the complexity:

Humanities

- Lines of transmissions;
- 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.

*"The fundamental methodological fact that **historical** linguists have to face is that they have no control over their data… The great art of the historical linguist is to make the **best of** this **bad data** - 'bad' in the sense that it may be **fragmentary**, **corrupted** or **many times removed** from the actual productions of native speakers."* (Labov, 1972, p. 100)

INTERTEXTUALITY

Mass digitisation is...

- driving the improvement of close reading methods;
- providing new opportunities for distant reading and for text reuse techniques to automatically find data parallels in large textual collections, i.e. intertextuality.

Jänicke, S., Franzini, G., et al. (2015) *Close and Distant Reading in Digital Humanities: A Survey and Future Challenges*. EuroVis Proceedings.

"[…] a text is […] a multidimensional space in which a variety of writings, none of them original, blend and clash. *The text is a tissue of quotations* drawn from the innumerable centres of culture… the writer can only imitate a gesture that is always anterior, never original. His only power is to mix writings […]." (Barthes, 1977, pp. 146-47)

"[…] any text is constructed as a *mosaic of quotations* […]." (Kristeva, 1980, p.66)

eTRAP's Aristotelian approach to intertextuality:

"The whole is greater than the sum of its parts"

Sources and socio-historical context or influencing factors (Gerard Genette's **epitext**).

CASE STUDY

DIGITAL BREADCRUMBS *of* **BROTHERS GRIMM**

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.

Motivation:

- Impact on society
- Global scope
- Big Data
- Interdisciplinary

Two avenues of enquiry:

1. **Intertextual relations between Grimm collection and other tales**;
2. Intertextual relations between Grimm collection and Grimm network.

Motif Database: to compare *Kinder- und Hausmärchen* to other tales by investigating measurable primitives:

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

Motif: ''[...] minimal thematic unit'' (Prince's *Dictionary of Narratology*)

Why build the database?

- Investigate & record primitives and their changes;
- Nothing like it exists;
- Advance research in folkloristics;
- Improve algorithms to sharpen our understanding of why and how a text is reused.

Selection: *Snow White, Puss in Boots, The Fisherman and his Wife*.

Two stages:

1. Manual
   - Collection of motifs as text reuse training data;
   - Creation of an ontology for RDF data representation.
2. (Semi-)Automatic
   - Detection of motifs in smaller corpora: TRACER;
   - Detection of motifs in larger corpora: Google Search & Books APIs.

Collection of motifs as text reuse training data

Reference: Aarne-Thompson Motif Index (1910; AT-Uther 2004).

Figure 2: Snow White motif collection file.

Creation of an ontology for RDF (Resource Description Framework) data representation

RDF Graph Database = VIRTUOSO
Query Language = SPARQL (RDF query language)

Output: a searchable database crossing the language barrier.

Figure 3: Mockup of user interface.

- Detection of motifs in smaller corpora: TRACER;
- Detection of motifs in larger corpora: Google Search & Books APIs.

TRACER

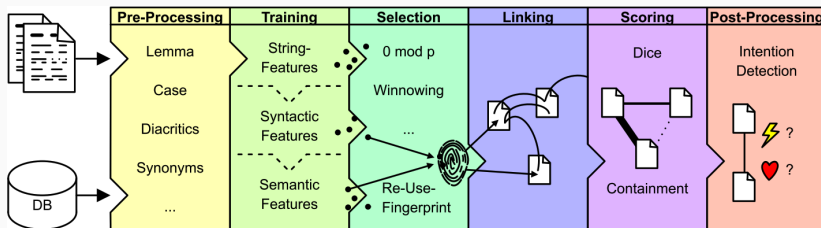TRACER: suite of 700 algorithms; developed by Marco Büchler.

Figure 4: TRACER steps. More than 1M permutations of implementations of different levels are possible.
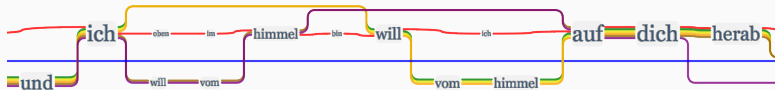
Figure 5: TRAViz display of the 7 Grimm variants of an *Aschenputtel* (= Cinderella) extract.

Text normalisation.

Jänicke, S., Geßner, A., Franzini, G., Terras, M., Mahony, S., Scheuermann, G. (2015) 'TRAViz: A Visualization for Variant Graphs', *Digital Scholarship in the Humanities (Digital Humanities 2014 Special Issue)* [Online]. DOI: `10.1093/llc/fqv049`

Two avenues of enquiry:

1. Intertextual relations between Grimm collection and other tales;
2. **Intertextual relations between Grimm collection and Grimm network.**
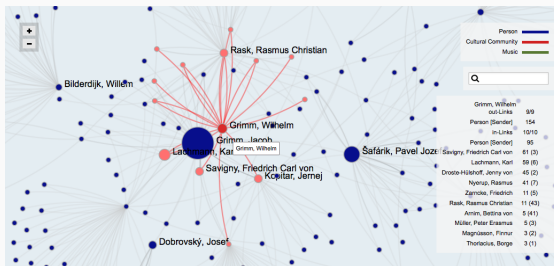
Figure 6: SPin's Grimm Cluster project: `http://ernie.spinnet.eu/viewer`
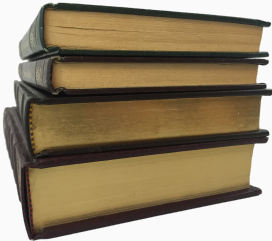
Figure 7: Grimm's book collection.
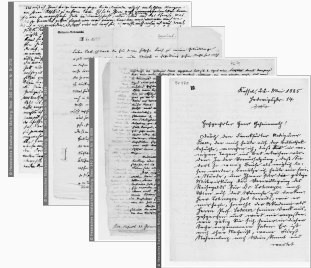
Figure 8: 36,000 Grimm letters.

CONCLUSION

- Contribution
  - Ontology of motifs
  - Motifs to Aarne-Thompson Motif-Index
- Next steps: short-term
  - Finish collecting motifs
  - Build the ontology
- Feedback
  - DH 2016 Kraków
  - You!

Author

Greta Franzini

✉ gfranzini@gcdh.de

Grimm Team (in alphabetical order)

Marco Büchler, Emily Franzini, Greta Franzini, Franziska Pannach, Gabriela Rotari, Christian Würker

Visit us

🌐 http://etrap.gcdh.de

"Copying from one is plagiarism, copying from many is research"
-Wilson Mitzner

Thank you!

REFERENCES

- Jänicke, S., Franzini, G., Faisal, C., Scheuermann, G. (2015) 'On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. A State-of-the-Art (STAR) Report', In: (Proceedings) *EuroVis 2015: The EG/VGTC Conference on Visualization*. Cagliari, May 2015, 25-29 [An extended version to be published in Eurographics Computer Graphics Forum (CGF)].

- Jänicke, S., Geßner, A., Franzini, G., Terras, M., Mahony, S., Scheuermann, G. (2015) 'TRAViz: A Visualization for Variant Graphs', *Digital Scholarship in the Humanities (Digital Humanities 2014 Special Issue)* [Online]. DOI: `10.1093/llc/fqv049`

- Labov, W. (1972) 'Some principles of linguistic methodology', *Language in Society*, 1(1), pp. 97-120 [Online]. At: `http://www.jstor.org/stable/4166672` (Accessed: 6 December 2015).

APPENDICES

⟨ðm∕⟩

Greta Franzini

Executive Board Member & Journal Associate Editor of ***Digital Medievalist***.

- Submit to or review for the Journal:
  `http://www.digitalmedievalist.org/journal/`

- Follow the website:
  `https://digitalmedievalist.wordpress.com/`

- Join the Facebook group: `https://goo.gl/Jr7EIG`

Questions & comments to: **gfranzini@gcdh.de**

All images used in this presentation are in the **public domain**.

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.