# DIGITAL BREADCRUMBS of BROTHERS GRIMM

Digital Humanities 2016
Kraków, 11-16 July

www.etrap.eu
Emily Franzini, Greta Franzini, Gabriela Rotari,
Franziska Pannach, Mahdi Solhdoust, Marco Büchler

## eTRAP

An early career research group focussing on the study of historical text reuse.
Text reuse describes the spoken and written repetition of content.

## RESEARCH QUESTIONS & OBJECTIVE

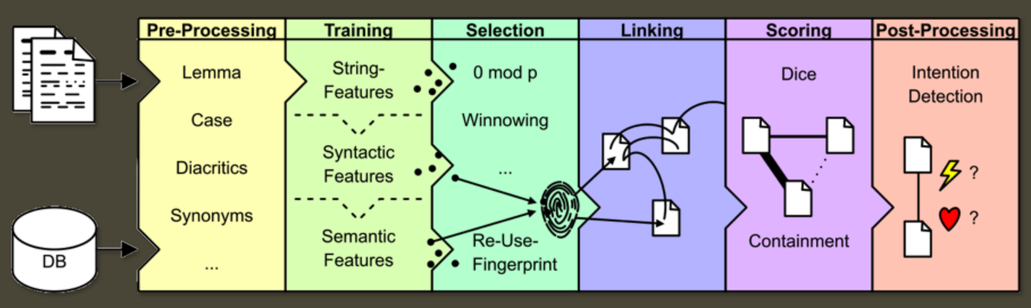
1. How does the human mind identify a reuse unit in context? Can machines do the same?
2. How can we detect text reuse across languages (big data)?
3. How reliable are online and offline retrieval systems in delivering stable result-sets when looking for reuse at scale?

Goal: The investigation of stability and volatility of text reuse and its primitives.
Objective: is to create a multilingual dataset in order to train text reuse algorithms to detect reuse units across languages and at scale.

## TRACER

eTRAP's TRACER is a suite of 700 algorithms, whose features can be combined to create the optimal formula for detecting those words, sentences and ideas that have been reused across texts. Specifically, the algorithms look for primitives, the stable elements of reuse units.

TRACER's framework comprises six steps.




## CURRENT DATASET: FAIRY TALES

1. Snow White (AT 709): Grimm [DE], Pushkin [RU], Tsvetaeva [RU], Calvino [IT], Jacobs [EN], Bruford [EN], Campbell [EN], Taylor [EN], Briggs [EN].
2. Puss in Boots (AT 545B): Grimm [DE], Straparola [IT], Pushkin [RU].
3. The Fisherman and his Wife (AT 555): Grimm [DE], Wild [DE], Pommern [DE], Pushkin [RU], Briggs [EN], Keding [EN], Andreev [UA].

The measurable primitives of reuse units in literature are motifs
Definition of motif "...minimal thematic unit[s]"
(Prince's Dictionary of Narratology)

## CASE STUDY: FAIRY TALE MOTIFS

Grimm's 'Schneewittchen' (Snow White) is comparable to Pushkin's 'Сказка о мертвой царевне и о семи богатырях' (The Tale of the Dead Princess and the Seven Knights). However, in the Grimm tale Snow White finds herself amongst seven dwarves, while in Pushkin's the young girl is protected by seven knights. In the Italian 'La Bella Venezia' (The beautiful Venice) by Calvino the girl is looked after by twelve thieves.

Despite the differences, the concept of 'the protected beautiful girl' links the tales.

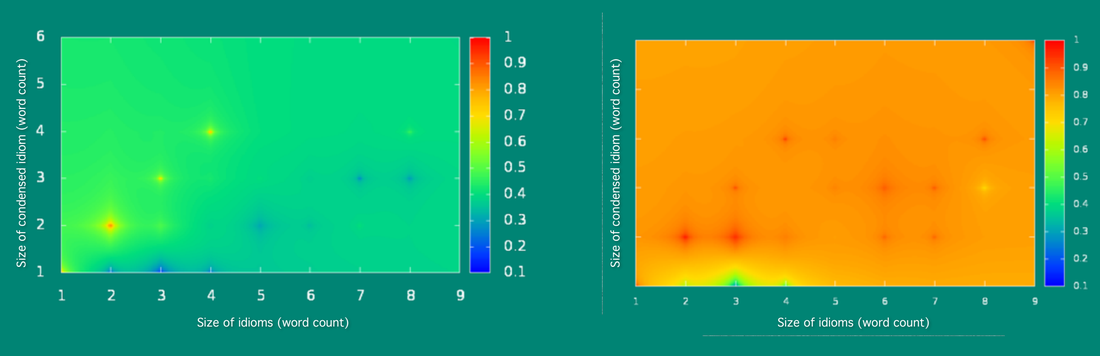
## METHOD

- We gather as many editions of the selected tales as possible in as many traditions/languages as we can. We exclude translations.
- Following the Aarne-Thompson (AT) index of fairy tale motifs we create a matrix in an Excel spreadsheet that confirms the presence or absence of a specific motif in the chosen tale and reduce motifs to keywords for processing.

| ISO Language Codes https://www.loc.gov/standards/ | GER |
|---|---|
| Aarne-Thompson | Grimm_1812 VIAF: 187449723 |
| D. MAGIC | |
| D800-D899. Ownership of magic objects | |
| D801. Ownership of magic object | nigin, haben, Spiegel, fragen, sprech |
| D812.6. Magic object received from witch or wizard | null |
| D900-D1299. Kinds of magic objects | |
| D930. Magic land features | null |
| D931. Magic rock (stone) | null |
| D931.0.5. Stone of patience | null |
| D1163. Magic mirror | Spiegel, sprechen |
| D1300-D1599. Function of magic objects | |
| D1310. Magic object gives supernatural information | te Frau im Land, Sneewittchen, taus |
| D1311. Magic object used for divination | n, über den sieben Bergen, tausendr |
| D1311.2. Mirror answers questions | |
| D1311.6. Divination by heavenly bodies | null |
| D1311.6.1. Moon (stars) answers questions | null |

- We integrate our multilingual data with the existing 'Thompson Motif Index' OWL/RDF ontology (by Koštová, Declerck & Klement).
- Our manually-gathered multilingual data is needed to refine text reuse detection algorithms used to trace motifs among larger and unstructured datasets.
- We trace reuse units at scale both online (Google Books and the web) and offline (Apache Lucene).

## TEXT REUSE AT SCALE




We conducted experiments with idioms as text reuse units. We used two approaches to identify these idioms at scale:
A. Google Custom Search (online): search for full and condensed version of idioms in Google Books and the web.
B. Apache Lucene (offline): search for full and condensed version of idioms in zeno.org (DARIAH-DE), Deutsches Textarchiv and Project Gutenberg.

Observations: the warmer the colour, the more similar the results. Lucene delivers significantly better results than Google when searching for full and condensed versions of idioms in text collections. This is because Google compromises on accuracy to favour speed, and because Lucene allows the user to set retrieval parameters for higher accuracy.

Bundesministerium für Bildung und Forschung
GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN
eTRAP Electronic Text Reuse Acquisition Project