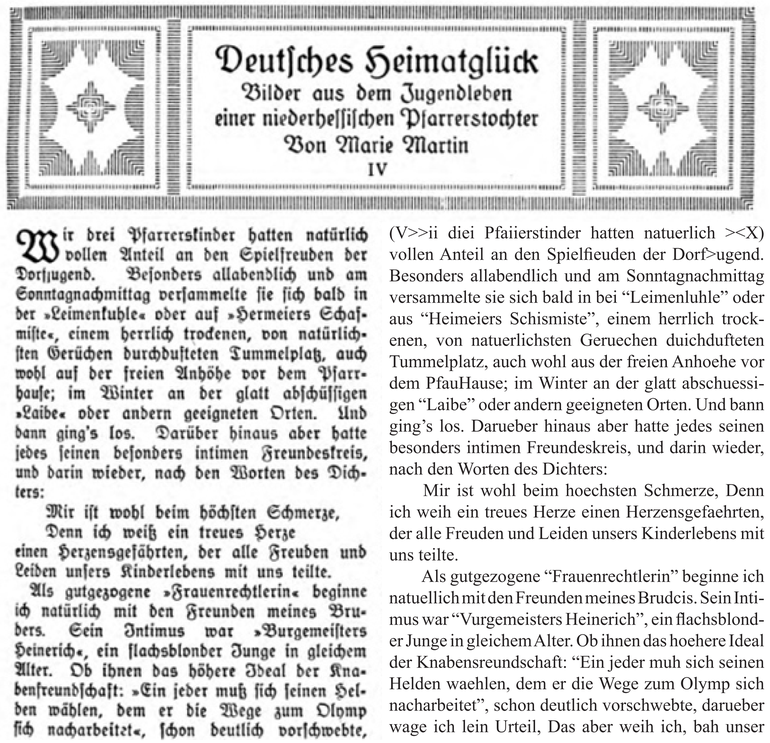
Digital Humanities & German Periodicals

Peter McIsaac, LSA/German *Sugih Jamin, EECS* *Ines Ibanez, SNRE*
Amy Blau, School of Information *Benjamin Bray, UROP*

Abstract

The unification of German speakers began well before formal unification in 1871 and continued into the twentieth century. The advent of mass-distributed periodicals accelerated this process, connecting geographically and politically separated German speakers with a stable source of news and entertainment. On an unprecedented scale, these periodicals captured the rapidly changing, slowly converging lifestyles of a national readership. Simultaneously, this enormous breadth precludes any exhaustive manual analysis of their contents, necessitating the development of intelligent automated heuristics to guide further research. In this project, we seek to extract salient contextual features from digitized versions of prevalent 19th-century periodicals using topic modelling. Due to digitization errors and unwanted, naturally-occurring artifacts, significant preprocessing is necessary before the digitized texts are useful. Various typographic characteristics of the printed texts contribute to a high rate of word and character misidentification, and unwanted features such as hyphenation, page numbers, and running headers must be reliably removed. First, we attempt to mend OCR errors using rule-based and stochastic techniques. Next, we repair word splits at line breaks and use fuzzy matching to remove headers and page numbers. Then, we use MALLET to estimate topics across our collection and assign a topic distribution to each document. Analysis of these distributions allows for efficient comparison and curation of large corpora, and may expose broad trends and interrelationships between documents. We hope to use this this automated "distant reading" to coarsely tag our corpus and guide a more scrupulous future investigation of developing 19th- and early 20th-century Germany.

Text Collections



Deutsches Heimatglück
Bilder aus dem Jugendleben
einer niederhessischen Pfarrerstochter
Von Marie Martin
IV

Wir drei Pfarrerskinder hatten natürlich vollen Anteil an den Spielfreuden der Dorfjugend. Besonders allabendlich und am Sonntagnachmittag versammelte sie sich bald in der »Leimenkuhle« oder auf »Hermeiers Schafmiſte«, einem herrlich trockenen, von natürlichſten Gerüchen durchdufteten Tummelplatz, auch wohl auf der freien Anhöhe vor dem Pfarrhauſe; im Winter an der glatt abſchüſſigen »Laibe« oder andern geeigneten Orten. Und dann ging's los. Darüber hinaus aber hatte jedes ſeinen beſonders intimen Freundeskreis, und darin wieder, nach den Worten des Dichters:

Mir iſt wohl beim höchſten Schmerze,
Denn ich weiß ein treues Herze
einen Herzensgefährten, der alle Freuden und Leiden unſers Kinderlebens mit uns teilte.

Als gutgezogene »Frauenrechtlerin« beginne ich natürlich mit den Freunden meines Bruders. Sein Intimus war »Burgemeiſters Heinerich«, ein flachsblonder Junge in gleichem Alter. Ob ihnen das höhere Ideal der Knabenfreundſchaft: »Ein jeder muß ſich ſeinen Helden wählen, dem er die Wege zum Olymp ſich nacharbeitet«, ſchon deutlich vorſchwebte,

(V>>ii diei Pfaiierstinder hatten natuerlich ><X) vollen Anteil an den Spielfieuden der Dorf>ugend. Besonders allabendlich und am Sonntagnachmittag versammelte sie sich bald in bei "Leimenluhle" oder aus "Heimeiers Schismiste", einem herrlich trockenen, von natuerlichsten Geruechen duichdufteten Tummelplatz, auch wohl aus der freien Anhoehe vor dem PfauHause; im Winter an der glatt abschuessigen "Laibe" oder andern geeigneten Orten. Und bann ging's los. Darueber hinaus aber hatte jedes seinen besonders intimen Freundeskreis, und darin wieder, nach den Worten des Dichters:

Mir ist wohl beim hoechsten Schmerze, Denn ich weih ein treues Herze einen Herzensgefaehrten, der alle Freuden und Leiden unsers Kinderlebens mit uns teilte.

Als gutgezogene "Frauenrechtlerin" beginne ich natuellich mit den Freunden meines Brudcis. Sein Intimus war "Vurgemeisters Heinerich", ein flachsblonder Junge in gleichem Alter. Ob ihnen das hoehere Ideal der Knabensreundschaft: "Ein jeder muh sich seinen Helden waehlen, dem er die Wege zum Olymp sich nacharbeitet", schon deutlich vorschwebte, darueber wage ich lein Urteil, Das aber weih ich, bah unser

Automatic OCR Correction

Fraktur

Texts from our German-language corpora were originally printed in the Fraktur typeface, leading to prohibitive OCR errors in their digitized form.

Aaä	Bb	Cc	Dd	Ee	Ff	Gg	Hh	Ii	Jj	Kk	Ll	Mm
a	b	c	d	e	f	g	h	i	j	k	l	m
Nn	Ooö	Pp	Qq	Rr	Sſs	Tt	Uuü	Vv	Ww	Xx	Yy	Zz
n	o	p	q	r	s	t	u	v	w	x	y	z

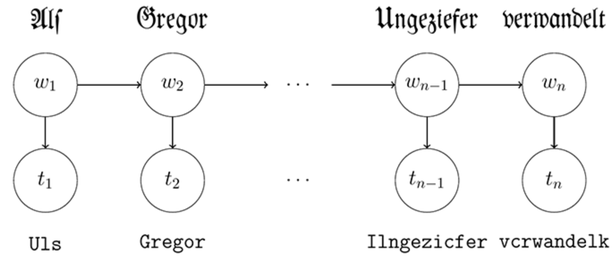
Noisy Channel Error Correction Scheme

We have begun development on a system to automatically locate and correct errors based on a **training corpus** of correct German text. We model the OCR process as a **noisy channel** that probabilistically alters correct input text, an approach shown to perform well with English.

For each sentence in the corpus, we perform the following steps:

- Flag the word as erroneous if it does not appear in our dictionary
- Generate lexically similar correction candidates
- Rank each candidate using a probabilistic measure and select the best one as a replacement for the erroneous word.

Correction candidates are selected based on the number of characters they share with the erroneous word. We use **Hidden Markov Model** to incorporate contextual information while ranking candidates in order to encourage the system to produce a corrected sequence of text that resembles valid German text.



We use our training corpus to estimate the probability that any given pair of German words (**bigram**) will appear together, encouraging uncommon word pairs in the corrected sequence.

Limitations & Future Work

Unfortunately, this approach to error correction performs very poorly on German text, likely as a result of the following factors:

- German words are highly **inflected** (adjective endings, etc.)
- German words are often combined to form **compounds**
- German is more **synthetic** than English (meaningful word fragments)

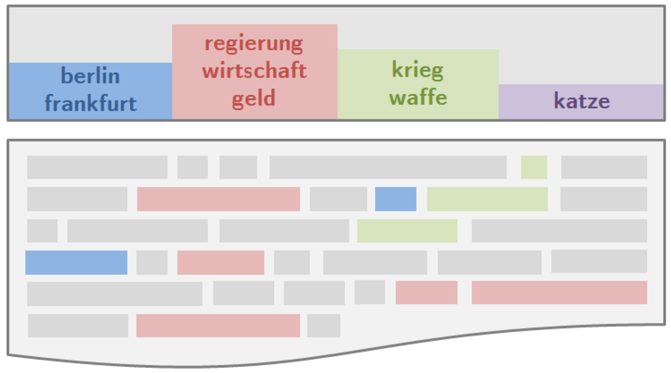
Each of these factors causes our system to conflate otherwise unrelated words, leading to poor correction choices. We have implemented a simple technique for splitting words into meaningful fragments (**morphs**). In the future, we may attempt to use the noisy channel correction scheme on sequences of morphs, rather than sequences of words, with the hope that this rough model of German morphology enables more informed correction decisions.

Topic Modeling & Analysis

Latent Dirichlet Allocation

Latent Dirichlet Allocation is a generative probabilstic model that can be used to extract groups of descriptive key words from a collection. Each document is divided into an unordered **bag of words** and analyzed with MALLET, an open-source topic modeling package. Under LDA, each **topic** defines a probability distribution over all possible German words, and each document is assumed to be generated in the following way:

1. For each document, assign random proportions to each topic
2. For each word,
   a. Choose a topic using the generated weights
   b. Choose a word according to the selected topic distribution



Visualization

We quantify the *similarity* between two documents by comparing their topic distributions and visualize the relationships between documents with a graph. This way, we can easily identify clusters of interesting documents to guide manual reading and expose broad relationships between documents that may not emerge in a manual search.