Integrating large-scale web data and curated corpus data in a search engine supporting German literacy education

Sabrina Dittrich\textsuperscript{a} \quad Zarah Weiss\textsuperscript{a}  
Hannes Schröter\textsuperscript{b} \quad Detmar Meurers\textsuperscript{a}

\textsuperscript{a} Department of Linguistics & LEAD Research Network  
University of Tübingen

\textsuperscript{b} German Institute for Adult Education – Leibniz Centre for Lifelong Learning, Bonn

8th Workshop on NLP for Computer Assisted Language Learning  
Turku, Finland, September 30th, 2019
A Search Engine for Literacy Education

- 30.6% of the German-speaking working age population have literacy skills below the level expected after 9th grade (Grotlüschen et al. 2019)
- Challenge to find reading materials for low literacy teaching
  - Lack of standardized didactic concepts
  - Few scientifically evaluated materials
  - Diverse biographic or educational learner background
- Search engine for text retrieval in literacy education
  - Retrieval of broad variety of high quality reading materials
  - Discrimination of readability levels at lowest literacy levels
Approaches to Text Retrieval

<table>
<thead>
<tr>
<th></th>
<th>Web Search</th>
<th>Curated Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Readability assessment</td>
<td>automatic</td>
<td>human</td>
</tr>
<tr>
<td>Up-to-date materials</td>
<td>✔</td>
<td>×</td>
</tr>
<tr>
<td>Broad text bandwidth</td>
<td>✔</td>
<td>×</td>
</tr>
<tr>
<td>Content quality control</td>
<td>×</td>
<td>✔</td>
</tr>
<tr>
<td>Materials at literacy level</td>
<td>×</td>
<td>✔</td>
</tr>
<tr>
<td>Clear copyright</td>
<td>×</td>
<td>✔</td>
</tr>
</tbody>
</table>


- Web search engines have **unique bandwidth** and are always **up-to-date**, but lack quality control and copyright.
- Systems using curated corpora contain texts of **unique quality**, but are limited in size and may become out-dated.

⇒ We combine both approaches for literacy education.
Characterizing Low Literacy in Germany

- Major studies in Germany that were supported by the Federal Ministry of Education and Research
  - *lea.* – *literacy development of workers* from 2008 to 2010
  - *leo.* – *Level-One* study
    (Grotlüschen & Riekmann 2011; Grotlüschen et al. 2019)

- Developed **ability-based descriptions for low literacy levels**
  - Alpha Level 1 to 3: literacy at letter, word or sentence-level
  - Alpha Level 4 to 6: increasing literacy at text-level

- We derived annotation guidelines and rule-based classification algorithm for **Alpha Readability Levels**
  (Weiss & Geppert 2018; Weiss, Dittrich & Meurers 2018)
Overview

- Leveled text retrieval for native- and non-native speakers of German in adult literacy and basic education classes

- Extends original *KANSAS Suche* (Weiss et al. 2018), which
  - allows to (de-)prioritizing linguistic constructions and re-rank retrieved materials accordingly.
  - automatically assigns Alpha Readability Levels to retrieved documents (Alpha Levels 3 to 6, and No Alpha).

- *KANSAS Suche 2.0* augments large-scale web search with curated materials for literacy education by providing
  - a filtered web search restricted to a pre-selected set of web providers of literacy education materials.
  - a corpus query in a curated corpus of materials for literacy education which we currently compile.
Workflow

**client**

- Enter Search term *either web or corpus search*

**server**

- **Web Search**
  - **Bing Search API**
  - **Text Extraction**
    - **Boilerpipe API**
  - **Linguistic Analysis**
    - **Stanford CoreNLP**
  - **Alpha Readability Level Classification**
    - *rule-based approach*
  - **Reranking, Filtering by Alpha Readability Level, Visualization**

- **Corpus Search**
  - **Solr Query**
  - **Loading of Preprocessed Documents**
    - **Deserialization of Java Objects**
      - (Text Extraction, Linguistic Analysis, Alpha Readability Classification has already been done offline)
    - **List of rankable documents**

**KANSAS Suche 2.0**

**System Description**
- Web Search Modes
- Corpus Search Mode
- Demonstration

**Comparison of Search Modes**
- Set-Up
- Coverage
- Readability
- Suitability
- Discussion

**Conclusion**

**References**

**Appendix**
Integrating web and corpus data in a search engine for literacy education

Sabrina Dittrich, Zarah Weiss, Hannes Schröter, and Detmar Meurers

Introduction

Text Retrieval Approaches
Low Literacy Skills

KANSAS Suche 2.0

System Description

Web Search Modes
Corpus Search Mode
Demonstration

Comparison of Search Modes

Set-Up
Coverage
Readability
Suitability
Discussion

Conclusion
References
Appendix
Web Search Mode

The unrestricted web search

- Queries use Microsoft Azure’s BING Web Search API

Web search on alpha sites

- Restricts web search to pre-compiled list of providers of reading materials for readers with low literacy skills
  - only 6 out of 75 reviewed web sites provided properly accessible contents for web engines

- Technically, the search restriction is implemented through standard site search operator.

⇒ 34,100 lexicon, news, and magazine texts in simple German, or written for children or language learners
Corpus Search Mode

The corpus search

▶ Query of pre-analyzed curated corpus using Apache Solr

The corpus

▶ High-quality materials for literacy education are scarce and those available often have unclear copyright

▶ We are compiling a corpus of high-quality materials of open educational resources with a corresponding license.

⇒ Collaboration with institutions that create materials for literacy and basic education
Demonstration of KANSAS Suche 2.0

KANSAS Suche 2.0: Search Modes
Comparison of Search Modes

- Web search access a large quantity of texts of poor readability and quality for low literate readers.
- The corpus search retrieves high-quality, readable texts but might fail to provide (enough) results for a query.
- Test assumption that search modes in KANSAS Suche 2.0 have complementing strengths and weaknesses.
- Compare search modes with regard to three criteria:
  - **Coverage** Returns requested number of results
  - **Readability** Results are readable for low literate readers
  - **Suitability** Results are suitable as teaching materials
Set-Up

- We queried with each search mode (www, corpus, filter) ten search terms requesting 30 results per term¹
  - As corpus, we used 10,012 semi-automatically cleaned texts crawled from web sites for low literate readers. (Weiss, Dittrich & Meurers 2018)
  - Search terms sampled from subset of basic vocabulary list for literacy education (Bockrath & Hubertus 2014)
  - We classified each of the texts using our readability classifier for Alpha Readability Levels (Weiss et al. 2018)
- We extracted a stratified sample for suitability annotation.

¹Alkohol (alcohol), Deutschkurs (German course), Erkältung (common cold), Heimat (home(land)), Internet (internet), Kirche (church), Liebe (love), Polizei (police), Radio (radio), and Staat (state)
Coverage across Search Modes

- We obtain 817 of 900 requested results
- Only the web search has full coverage
- Corpus and filtered search struggle with only two terms
Readability across Search Modes

<table>
<thead>
<tr>
<th></th>
<th>WWW + length</th>
<th>WWW - length</th>
<th>Filter + length</th>
<th>Filter - length</th>
<th>Corpus + length</th>
<th>Corpus - length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha 3</td>
<td>0.00%</td>
<td>1.00%</td>
<td>0.39%</td>
<td>4.30%</td>
<td>4.98%</td>
<td>13.41%</td>
</tr>
<tr>
<td>Alpha 4</td>
<td>19.00%</td>
<td>49.67%</td>
<td>15.23%</td>
<td>53.91%</td>
<td>35.25%</td>
<td>50.19%</td>
</tr>
<tr>
<td>Alpha 5</td>
<td>14.33%</td>
<td>21.00%</td>
<td>8.20%</td>
<td>22.66%</td>
<td>14.56%</td>
<td>18.77%</td>
</tr>
<tr>
<td>Alpha 6</td>
<td>10.00%</td>
<td>2.00%</td>
<td>7.42%</td>
<td>10.16%</td>
<td>8.43%</td>
<td>7.28%</td>
</tr>
<tr>
<td>No Alpha</td>
<td>56.67%</td>
<td>26.33%</td>
<td>68.75%</td>
<td>8.98%</td>
<td>36.78%</td>
<td>10.34%</td>
</tr>
</tbody>
</table>

- Web search retrieves least Alpha 3 and many No Alpha texts, but surprising number of Alpha 4 texts.
- Corpus search has few No Alpha and most Alpha 3 texts.
- Filtered search retrieves least No Alpha texts (excl. length), but less than 5% Alpha 3 texts → middle ground.
Suitability

- Annotate sample of texts stratified by readability levels, search terms and search modes ($N = 451$)

- Annotation of ±suitable by two annotators ($\kappa = 0.77$), with guidelines specifying as not suitable:
  - advertisement, brief captions of graphics, and hubs
  - pages not containing the search term or a synonym
  - texts with more than 1,500 words

⇒ Consider texts are not suitable if both annotators agree
  - Overall 30.38% of the sample are not suitable (137 texts)
Suitability across Search Modes and Terms

- Web search highest rate of not suitable materials (52.70%), filtered search has 31.00%, and corpus search 8.80%

- *Deutschkurs* (German course) but also *Internet* and *Radio* elicit more not suitable results than others

- Unlike both web searches, the corpus search hardly elicits materials that are not suitable.
Suitability and Readability across Search Modes

▶ Majority of Alpha 4 material retrieved by either web search is not suitable.

▶ Many Alpha 5 and nearly all Alpha 6 materials retrieved by the unrestricted search are unsuitable.

▶ For nearly all Alpha Levels the corpus search has better coverage for suitable and readable material.
Discussion

▶ The unrestricted web search in principle has broader coverage, but finds less readable and suitable materials.

▶ The restricted web search and the corpus search yield satisfying coverage, especially after considering suitability.

▶ The best search mode is dependent on the search goal:
  ▶ The rate of not suitable results in web search and coverage in other search modes are search term dependent.
  ▶ A corpus search is best to find Alpha 3 texts.
  ▶ The filtered search compromises between web and corpus.
Conclusion & Outlook

▶ We combine the strengths of web search and corpus data for text retrieval of literacy education materials.

▶ There is no universally best search approach which makes flexible choices between search modes important.

▶ The system is fully implemented and will be officially released upon completion of the curated corpus.

▶ We plan to expand the corpus search functionality to also support a fully linguistic retrieval without content search.
References


The Filtered Web Search

- Simple German: nachrichtenleicht.de, hurraki.de/wiki, lebenshilfe.de/de/leichte-sprache
- For Children: klexikon.zum.de, geo.de/geolino
- For language learning: deutsch-perfekt.com
## Readability Algorithm

<table>
<thead>
<tr>
<th>Feature</th>
<th>Alpha 3</th>
<th>Alpha 4</th>
<th>Alpha 5</th>
<th>Alpha 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>W / S</td>
<td>10</td>
<td>10</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Syllables / W</td>
<td>3</td>
<td>5</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Dep. clauses / S</td>
<td>≤0.50</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Unknown voc.</td>
<td>≥0.95</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Lat./Greek Nouns</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Present tense</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Simple past</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Perfect</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Future</td>
<td>X</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Past Perfect</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>✓</td>
</tr>
</tbody>
</table>

S = sentence; W = word
The Test Corpus

- Texts from more than 25 projects by over 20 organizations
- Contains expositions, glosses, narratives, articles, podcasts, recipes, and wikis
- Includes texts for different target groups
  - 6,341 texts in simple German
  - 3,022 in simplified German
  - 649 by people with low literacy skills
Readability across Search Modes and Terms

- Few notable effects of search term on text readability
- Less Alpha 4 texts for Erkältung (common cold) and Deutschkurs (German lesson) (low coverage)
- Overall, readability is comparable across search terms
  ➞ Expected, since all terms are from a basic vocabulary list

<table>
<thead>
<tr>
<th>Search Term</th>
<th>Alpha 3</th>
<th>Alpha 4</th>
<th>Alpha 5</th>
<th>Alpha 6</th>
<th>no Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staat (state)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Radio (radio)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polizei (police)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liebe (love)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kirche (church)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet (internet)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heimat (home(land))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Erkältung (common cold)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deutschkurs (German lesson)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alkohol (alcohol)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Search Term</th>
<th>Alpha 3</th>
<th>Alpha 4</th>
<th>Alpha 5</th>
<th>Alpha 6</th>
<th>no Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 20 40 0 20 40 0 20 40 0 20 40 0 20 40 0 20 40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alkohol (alcohol)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deutschkurs (German lesson)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Erkältung (common cold)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>