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

Sabrina Dittrich[a] Zarah Weiss[a]
Hannes Schröter[b] Detmar Meurers[a]

[a] Department of Linguistics & LEAD Research Network
University of Tübingen

[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

$\Rightarrow$ 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

	Web Search	Curated Corpus
Readability assessment	automatic	human
Up-to-date materials	✓	✗
Broad text bandwidth	✓	✗
Content quality control	✗	✓
Materials at literacy level	✗	✓
Clear copyright	✗	✓

- 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.

$\Rightarrow$ 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

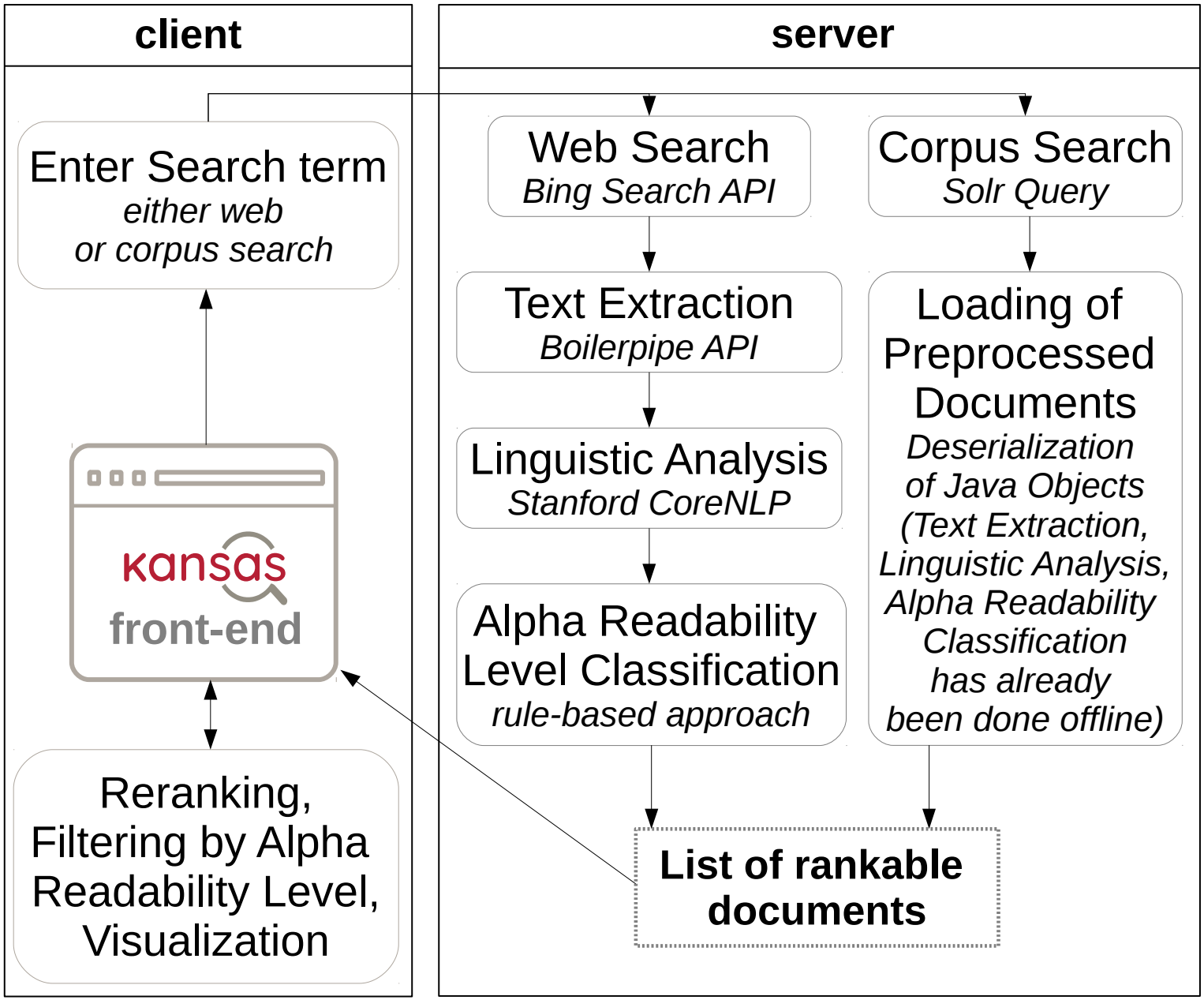
  (Grotlüschen & Riekmann 2011; Grotlüschen et al. 2019)
- Developed ability-based descriptions for low literacy levels
  - Alpha Level 1 to 3: literacy at letter, word or sentence-level
  - Alpha Level 4 to 6: increasing literacy at text-level
- We derived annotation guidelines and rule-based classification algorithm for Alpha Readability Levels

  (Weiss & Geppert 2018; Weiss, Dittrich & Meurers 2018)

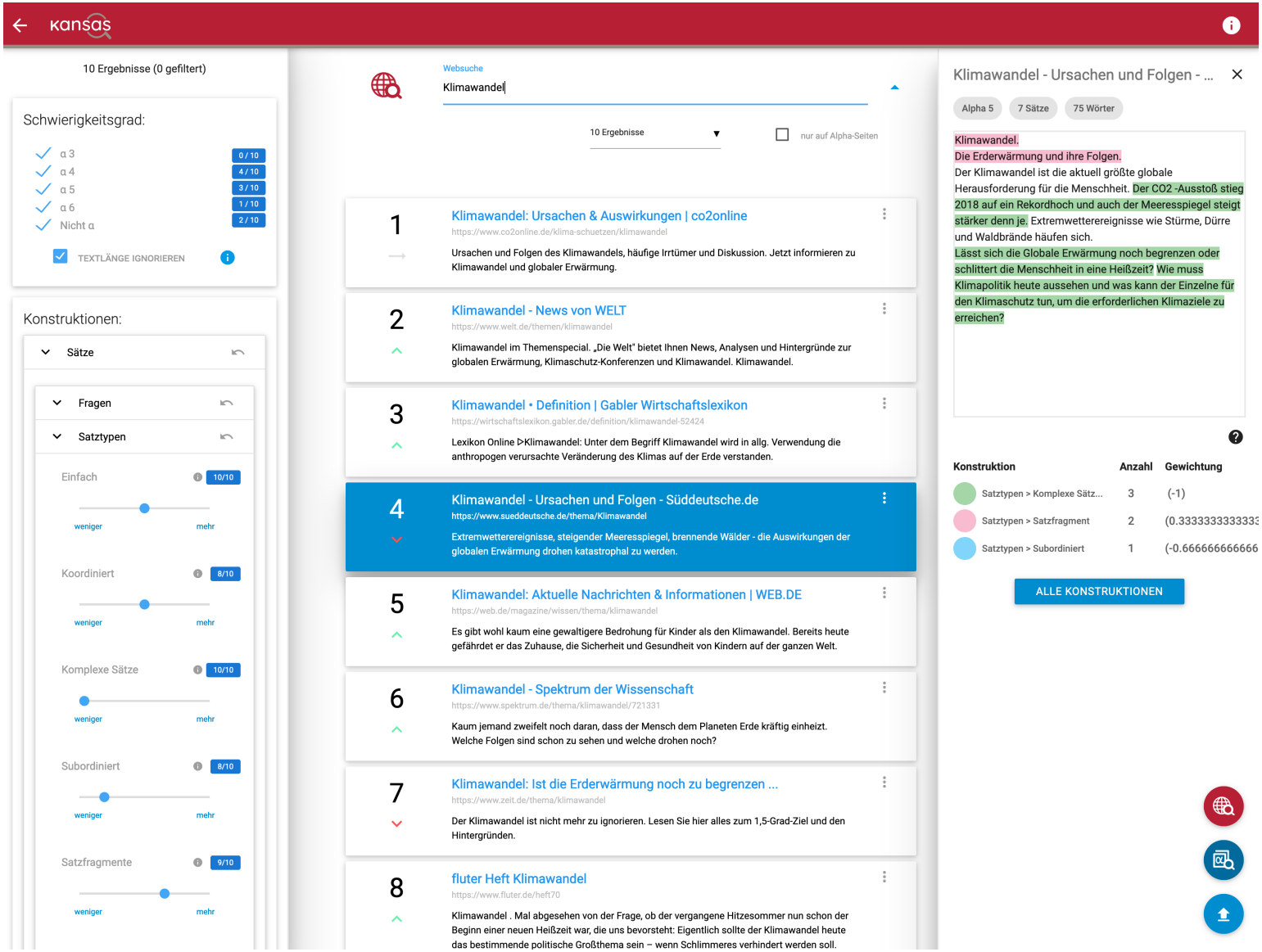
Overview

- 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



KANSAS Suche Result View



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.

$\Rightarrow$ 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.

$\Rightarrow$ 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

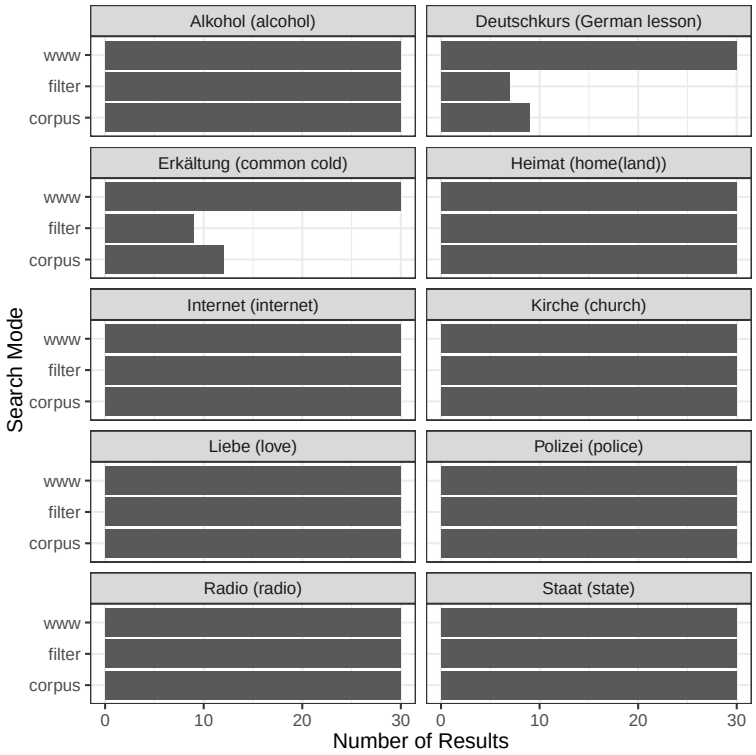
Coverage	Returns requested number of results
Readability	Results are readable for low literate readers
Suitability	Results are suitable as teaching materials

Set-Up

- We queried with each search mode (www, corpus, filter) ten search terms requesting 30 results per term[1]
  - 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.

[1] Alkohol (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

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

- 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

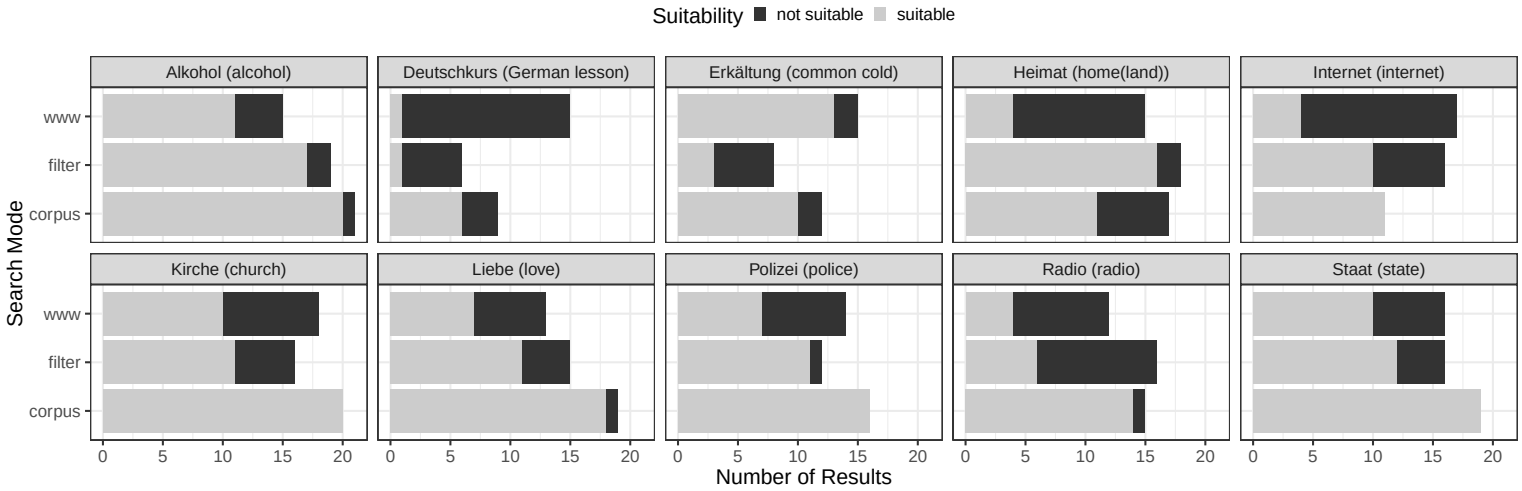
Suitability

- 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 specifiying 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

$\Rightarrow$ Consider texts are not suitable if both annotators agree

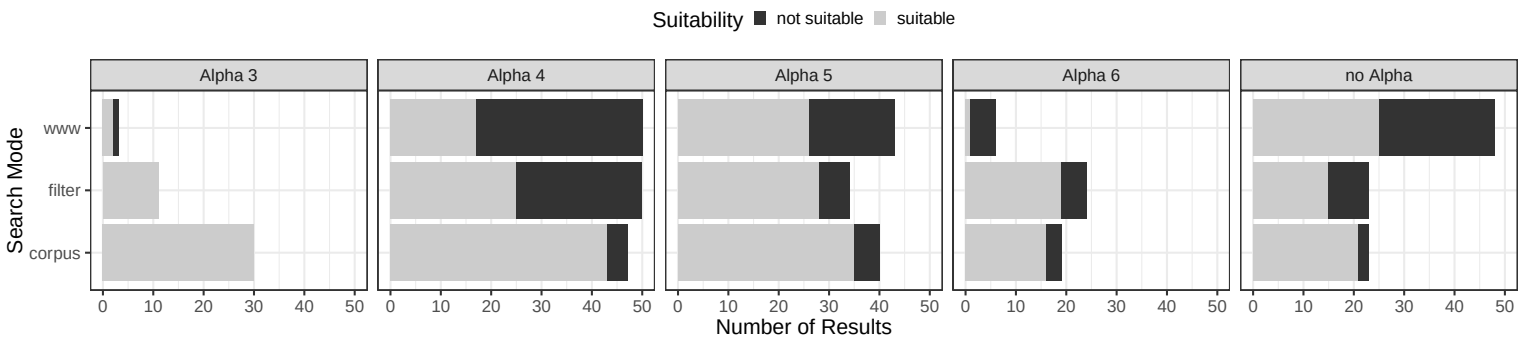
- Overall 30.38% of the sample are not suitable (137 texts)

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

Bockrath, A. & P. Hubertus (2014). *1.300 wichtige Wörter. Ein Grundwortschatz*. Münster, Germany: Bundesverband Alphabetisierung und Grundbildung e.V., 5th ed.

Grotlüschen, A., K. Buddeberg, G. Dutz, L. Heilmann & C. Stammer (2019). LEO 2018 – living with low literacy. Press brochure, Hamburg, Germany. http://blogs.epb.uni-hamburg.de/leo/files/2019/06/LEO_2018_Living_with_Low_Literacy.pdf.

Grotlüschen, A. & W. Riekmann (2011). leo. - Level-Online Studie. Press brochure, Hamburg, Germany. http://blogs.epb.uni-hamburg.de/leo/files/2011/12/leo-Press-brochure15-12-2011.pdf.

Weiss, Z., S. Dittrich & D. Meurers (2018). A Linguistically-Informed Search Engine to Identifiy Reading Material for Functional Illiteracy Classes. In *Proceedings of the 7th Workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL)*. Association for Computational Linguistics.

Weiss, Z. & T. Geppert (2018). Textlesbarkeit für Alpha-Levels. Annotationsrichtlinien für Lesetexte. Version 1.1. http://www.sfs.uni-tuebingen.de/~zweiss/rsrc/textlesbarkeit-fur-alpha.pdf.

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

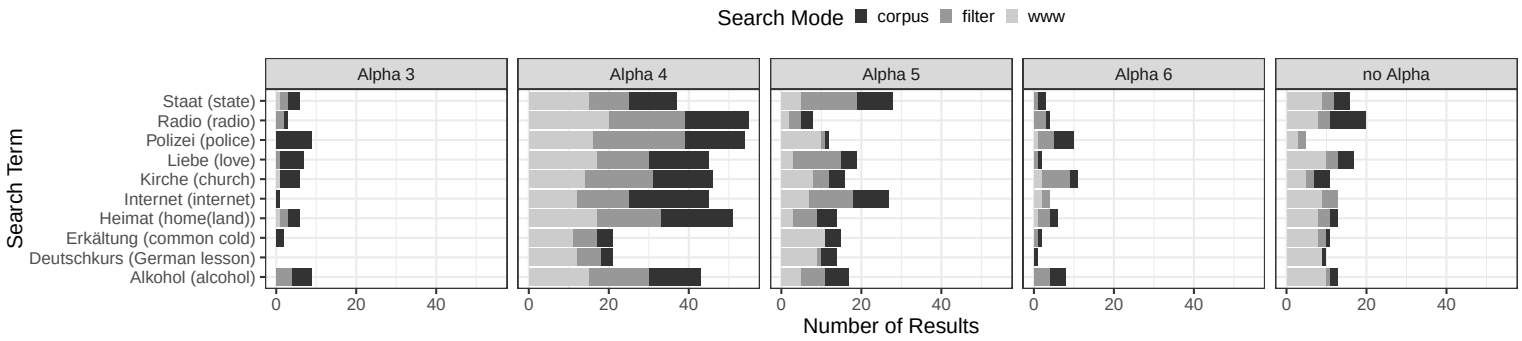
Feature	Alpha 3	Alpha 4	Alpha 5	Alpha 6
W / S	10	10	12	12
Syllables / W	3	5	✓	✓
Dep. clauses / S	≤0.50	✓	✓	✓
Unknown voc.	≥0.95	✓	✓	✓
Lat./Greek Nouns	✗	✓	✓	✓
Present tense	✓	✓	✓	✓
Simple past	✗	✓	✓	✓
Perfect	✗	✓	✓	✓
Future	✗	✗	✓	✓
Past Perfect	✗	✗	✗	✓

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

$\Rightarrow$ Expected, since all terms are from a basic vocabulary list