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**[c] **Detmar Meurers**[a,b]
[a] Department of Linguistics, University of Tübingen
[b] LEAD Graduate School and Research Network, University of Tübingen
[c] German Institute for Adult Education – Leibniz Centre for Lifelong Learning
{dittrich,zweiss,dm}@sfs.uni-tuebingen.de schroeter@die-bonn.de

Abstract

Reading material that is of interest and at the right level for learners is an essential component of effective language education. The web has long been identified as a valuable source of reading material due to the abundance and variability of materials it offers and its broad range of attractive and current topics. Yet, the web as source of reading material can be problematic in low literacy contexts.

We present ongoing work on a hybrid approach to text retrieval that combines the strengths of web search with retrieval from a high-quality, curated corpus resource. Our system, *KANSAS Suche 2.0*, supports retrieval and reranking based on criteria relevant for language learning in three different search modes: unrestricted web search, filtered web search, and corpus search. We demonstrate their complementary strengths and weaknesses with regard to coverage, readability, and suitability of the retrieved material for adult literacy and basic education. We show that their combination results in a very versatile and suitable text retrieval approach for education in the language arts.

1 Introduction

Low literacy skills are an important challenge for modern societies. In Germany, 12.1% of the German-speaking working age population (18 to 64 years), approximately 6.2 million people, cannot read and write even short coherent texts; another 20.5% cannot read or write coherent texts of medium length (Grotlüschen et al., 2019), falling short of the literacy rate expected after nine years of schooling. While these figures are lower than those reported in previous years (Grotlüschen and Riekmann, 2011), there seems to be no significant change in the proportion of adults with low literacy skills when taking into account demographic changes in the composition of the population from 2010 to 2018 (Grotlüschen and Solga, 2019, p. 34), such as the risen employment rate and average level of education.

Literacy skills at such a low level impair the ability to live independently, to participate freely in society, and to compete in the job market. To address this issue, the German federal and state governments launched the *National Decade for Literacy and Basic Skills (AlphaDekade) 2016–2026*.[1] One major concern in the efforts to promote literacy and basic education in Germany is the support of teachers of adult literacy and basic education classes, who face particular challenges in ensuring the learning success of their students. Content-wise, teaching materials should be of personal interest to them and closely aligned with the demands learners face in their everyday and working life (BMBF and KMK, 2016, p. 6). Language-wise, reading materials for low literacy and for language learning in general should be authentic (Gilmore, 2007) and match the individual reading skills of the learners so that they are challenged but not overtaxed (Krashen, 1985; Swain, 1985; Gilmore, 2007). This demand for authentic, high-quality learning materials is currently not met by publishers, making it difficult for teachers to address the needs of their diverse literacy classes. Relatedly, there is a lack of standardized didactic concepts and scientifically evaluated materials, despite first efforts to address this shortage (Löffler and Weis, 2016). What complicates matters fur-

[1] https://www.alphadekade.de

ther is that literacy and basic education classes are comprised of learners with highly heterogeneous biographic and education backgrounds. This includes native and non-native speakers, the latter of whom may or may not be literate in their native language. Low literacy skills sometimes are also associated with neuro-atypicalities such as intellectual disorders, dyslexia, or Autism Spectrum Disorders (ASD; Friedman and Bryen, 2007; Huenerfauth et al., 2009).

Given the shortage of appropriate reading materials provided by publishers, the web is an attractive alternative source for teachers seeking reading materials for their literacy and basic education classes. There is an exceptional coverage of current topics on the web, and a standard web search engine provides English texts at a broad range of reading levels (Vajjala and Meurers, 2013), though the average reading level of the texts is quite high. For German, most online reading materials appear to target native speakers with a medium to high level of literacy. Offers for low literate readers are restricted to a few specialized web pages presenting information in simple language or simplified language for language learners or children. These may or may not be suited in content and presentation style for low literate adults. Web materials specifically designed for literacy and basic education do not follow a general standard indicating how the appropriateness of the material for this context was assessed. This makes it difficult for teachers of adult literacy and basic education classes to verify the suitability of the materials. As we argued in Weiss et al. (2018), this challenge extends beyond the narrow context of literacy classes, as it also pertains to the question of web accessibility for low literate readers who perform their own web queries.

We address this issue by presenting our ongoing work on *KANSAS Suche 2.0*, a hybrid search engine for low literacy contexts that offers three search modes: free web search, filtered web search exclusively on domains providing reading materials for low levels of literacy (henceforth: *alpha sites*), and a corpus of curated, high-quality literacy and basic education materials we are currently compiling. The corpus will come with a copyright allowing teachers to adjust and distribute the materials for their classes. We thus considerably extend the original *KANSAS Suche* system (Weiss et al., 2018), which only supported web search. Different from previous text retrieval systems for language learning, focusing on either web search or compiled text repositories (Heilman et al., 2010; Collins-Thompson et al., 2011; Walmsley, 2015; Chinkina et al., 2016), our approach instantiates a hybrid architecture in the spectrum of potential strategies (Chinkina and Meurers, 2016, Figure 4) by combining the strengths of focused, high-quality text databases with large-scale, more or less parameterized web search.

The remainder of the article is structured as follows: First, we briefly review research on readability and low literacy and compare previous approaches to text retrieval systems for education contexts (section 2). Then, we describe our system in section 3, before providing a quantitative and qualitative comparison of the three different search modes supported by our system in section 4. Section 5 closes with some final remarks on future work.

2 Related Work

Text retrieval for low literate readers or language learners at its core consists of two tasks: text retrieval, and readability assessment of the retrieved texts. We here provide some background on previous work on these two tasks as well as on the German debate on how to characterize low literacy skills. We start by reviewing work on readability analysis for language learning and low literacy contexts (section 2.1), before discussing the characterization of low literacy skills (section 2.2), and ending with an overview of text retrieval approaches for language learning (section 2.3).

2.1 Readability Assessment

Automatic readability assessment matches texts to readers with a certain literacy skill such that they can fulfill a predefined reading goal or task such as extracting information from a text. Early work on readability assessment started with readability formulas (Kincaid et al., 1975; Chall and Dale, 1995) which are still used in some studies (Grootens-Wiegers et al., 2015; Esfahani et al., 2016) despite having been widely criticized for being too simplistic and unreliable (Feng et al., 2009; Benjamin, 2012). In answer to this criticism, more advanced methods supporting broader linguistic modeling using Natural Language Processing (NLP) were established. For example, Vajjala and Meurers (2012) showed that measures of language complexity originally devised in Second Language Acquisition (SLA) research can successfully be adopted to the task of readability classification. An increasing amount of NLP-based research is being dedicated to the assessment of readability for different contexts, in particular for English (Feng et al., 2010; Crossley et al., 2011; Xia et al., 2016; Chen and Meurers, 2017b), with much less work on other languages, such as French, German, Italian, and Swedish (François and Fairon, 2012; Weiss and Meurers, 2018; Dell'Orletta et al., 2011; Pilán et al., 2015).

Automatic approaches to readability assessment at low literacy levels are less common, arguably also due to the lack of labeled training data for the highly heterogeneous group of adults with low literacy in their native language (Yaneva et al., 2016b). But there is research in this domain bringing in eye-tracking evidence to identify challenges and reading strategies for neuro-atypical readers with low literacy skills, such as people with dyslexia (Rello et al., 2013a,b) or ASD (Yaneva et al., 2016a; Eraslan et al., 2017). Two approaches should be mentioned that overcome the lack of available training data by implementing rules determined in previously developed guidelines for low literacy contexts. Yaneva (2015) presents a binary classification approach to determine the adherence of texts to Easy-to-read guidelines. Easy-to-read guidelines are designed to promote accessibility of reading materials for readers with cognitive disabilities such as 'Make It Simple' by Freyhoff et al. (1998) and 'Guidelines for Easy-to-read Materials' by Nomura et al. (2010). Yaneva (2015) applies this algorithm to web materials labeled as Easy-to-Read to investigate their compliance to the guidelines by Freyhoff et al. (1998). She shows that providers of Easy-to-Read materials overall adhere to the guidelines. This is an important finding since not all self-declared 'simple' reading materials on the web actually are suitable for readers with lower reading skills. For example, Simple Wikipedia was found to not be systematically simpler than Wikipedia (see, for example, Štajner et al., 2012, Xu et al., 2015, and Yaneva et al., 2016b), though Vajjala and Meurers (2014) illustrate that an analysis at the sentence level can identify relative complexity differences. While such research on the adherence of web materials to guidelines is an important contribution to the evaluation of web accessibility, it is less suitable for our education purposes, as it does not differentiate degrees of readability within the range of low literacy. In Weiss et al. (2018), we propose a rule-based classification approach for German following the analysis of texts for low literate readers in terms of the so-called Alpha Readability Levels introduced in the next section. As far as we are aware, this currently is the only automatic readability classification approach that differentiates degrees of readability at such low literacy levels.

2.2 Characterizing Low Literacy Skills

According to recent large-scale studies, there is a high proportion of low literate readers in all age groups of the population in Germany (Schröter and Bar-Kochva, 2019). For the German working age population (18–64 years), three major studies further focused on investigating the parts of the working population with the lowest level of literacy skills. The *lea. – Literalitätsentwicklung von Arbeitskräften* [literacy development of workers] study carried out from 2008 to 2010, supported by the Federal Ministry of Education and Research, was the first national survey on reading and writing competencies in the German adult population.[2] In this context, a scale of six *Alpha Levels* was developed to allow a fine grained measure of the lowest levels of literacy. These levels were empirically tested in the first *leo. – Level-One* study (Grotlüschen and Riekmann, 2011), which was updated in 2018 (Grotlüschen et al., 2019).[3]

At Alpha Level 1, literacy is restricted to the level of individual letters and does not extend to the word level. At Alpha Level 2, literacy is restricted to individual words and does not extend to the sentence level. At Alpha Level 3, literacy is restricted to single sentences and does not extend to the text level. At Alpha Levels 4 to 6, literacy skills are sufficient to read and write increasingly longer and more complex texts. The descriptions of Alpha Levels in the *lea.* and *leo.* studies are ability-based, i.e., they focus on what someone at this literacy level can and cannot read or write. Weiss and Geppert (2018) used these descriptions to derive annotation guidelines for the assessment of texts rather than people, focusing on the Levels 3 to 6 relevant for characterizing texts. Based on those annotation guidelines, Weiss et al. (2018)

[2] `https://blogs.epb.uni-hamburg.de/lea`

[3] `https://blogs.epb.uni-hamburg.de/leo`

developed a rule-based algorithm supporting the automatic classification of reading materials for low literacy contexts. They demonstrate that the classifier successfully approximates human judgments of Alpha Readability Levels as operationalized in Weiss and Geppert (2018).

While Alpha Levels 4 and higher still describe very low literacy skills, only Alpha Levels 1 to 3 constitute what has previously been referred to as *functional illiteracy* in the German adult literacy discourse: literacy skills at a level that only permits reading and writing below the text level. Literacy at this level is not sufficient to fulfill standard social requirements regarding written communication in the different domains of working and living (Decroll, 1981). Grotlüschen et al. (2019) argue that the term functional illiteracy is stigmatizing and therefore ill-suited for use in adult education. Instead, they refer to *adults with low literacy skills* and *low literacy*. In the following, we will use the term *low literacy* in a broad sense to refer to literacy skills up to and including Alpha Level 6. To discuss literacy below the text level, i.e., Alpha Levels 1 to 3, we will make this explicit by referring to *low literacy in a narrow sense*.

2.3 Text Retrieval for Language Learning

A growing body of research is dedicated to the development of educational applications that provide reading materials for language and literacy acquisition. Many of them are forms of leveled search engines, i.e., information retrieval systems that perform some form of content-based web query and analyze the readability of the retrieved materials on the fly, often by using readability formulas as discussed in section 2.1. The readability level of the results is then displayed to the user as additional criterion for the text choice, the results are ranked according to the level, or a readability filter allows exclusion of hits with undesired readability levels. The majority of these systems are designed for English (Miltsakaki and Troutt, 2007; Collins-Thompson et al., 2011; Chinkina et al., 2016), although there are some notable exceptions for a few other languages (Nilsson and Borin, 2002; Walmsley, 2015; Weiss et al., 2018). One of the main advantages of leveled web search engines is that they allow access to a broad bandwidth of texts that are always up-to-date. These are important features for the identification of interesting and relevant reading materials in educational contexts. Beyond the educational domain, leveled web search engines also contribute to web accessibility by allowing web users with low literacy skills to query web sites that are at a suitable reading level for their purposes. One example for such a system for literacy training is the original *KANSAS Suche* (Weiss et al., 2018). The system analyses web search results and assigns reading levels to them based on a rule-based algorithm which is specifically designed for low literate readers. Linguistic constructions can be (de-)prioritized to re-rank the search results.

The main drawback of such web-based approaches, however, is the lack of control of the quality of the content. This may lead to results that include incorrect or biased information or inappropriate materials, such as propaganda, racist or discriminating contents, or fake news. This issue may require the attentive eye of a teacher during the selection process. Query results may also include picture or video captions, forum threads, or shopping pages, which are unsuited as reading texts. To avoid such issues, many applications rely on restricted web searches on pre-defined websites, as is the case for FERN (Walmsley, 2015) or net-Trekker (Huff, 2008), which may also be crawled and analyzed beforehand, as in SyB (Chen and Meurers, 2017a).

Some systems extend their functionality beyond a leveled search engine and incorporate tutoring system functions. For example, the FERN search engine for Spanish (Walmsley, 2015) provides an enhanced reading support by allowing readers to flag and look up unknown words and train new vocabulary in automatically generated training material. This relates to another type of educational application that provides reading materials for language and literacy acquisition: reading tutors. Such systems generally provide access to a collection of texts that have been collected and analyzed beforehand (Brown and Eskenazi, 2004; Heilman et al., 2008; Johnson et al., 2017). The collections are usually a curated selection of high-quality texts that are tailored towards the specific needs of the intended target group. To function as tutoring systems, the systems support interaction for specific tasks, e.g., reading comprehension or summarizing tasks. One example for such a system in the domain of literacy and basic education is iSTART-ALL, the Interactive Strategy Training for Active Reading and Thinking for Adult Literacy Learners (Johnson et al., 2017). It is an intelligent tutoring system for reading comprehension with several practice modules, a text library, and an interactive narrative. It contains a set of 60 simplified news stories sampled from the California Distance Learning Project.[4] They are specifically designed to address the interests and needs of adults with low literacy skills (technology, health, family). It offers summarizing and question asking training for these texts as well as an interactive narrative with integrated tasks and immediate corrective feedback. The greater quality of curated reading materials in reading tutors comes at the cost of drawing from a considerably more limited pool of reading materials, which may become obsolete quickly. Thus, leveled web search engines as well as reading tutors have complementary strengths and weaknesses. As we will demonstrate in the following, combining the two approaches can help obtain the best of both worlds.

3 Our System

We present *KANSAS Suche 2.0*, a hybrid search system for the retrieval of appropriate German reading materials for literacy and basic education classes. While these classes are typically designed for low literate native speakers, in practice, they are comprised of native- and non-native speakers of German.[5] As the original *KANSAS Suche* (Weiss et al., 2018), which was inspired by *FLAIR* (Chinkina and Meurers, 2016; Chinkina et al., 2016), the updated system operates on the premise that users want to select reading materials in a way combining content queries with a specification of the linguistic forms that should be richly represented or not be included in the text. But *KANSAS Suche 2.0* is a hybrid system in the sense that it combines different search modes in order to overcome the individual weaknesses of web-based and corpus-based text retrieval outlined in the previous section.

More specifically, our system offers three different search modes: a) an unrestricted web search option to perform large-scale content queries on the web, b) a filtered web search to perform content queries on web pages specifically designed for low literacy and basic education purposes, and c) a corpus search mode to retrieve edited materials that have been pre-compiled specifically for the purpose of literacy and basic education courses. Users may flexibly switch between search modes if they find that for a specific search term the chosen search mode does not yield results satisfying their needs. In all three search modes, the new system allows users to re-rank search results based on the (de-)prioritization of linguistic constructions, just as in the original *KANSAS Suche*. The results are automatically leveled by readability in terms of an Alpha Level-based readability scale specifically tailored towards the needs of low literacy contexts following Kretschmann and Wieken (2010) and Gausche et al. (2014), as detailed in Weiss et al. (2018). This readability classification may be used to further filter results, to align them with the reading competencies of the intended reader. Users can also upload their own corpora to re-rank the texts in them based on their linguistic characteristics and automatically compute their Alpha Readability Levels. We understand this upload functionality as an additional feature rather than a separate search mode because it does not provide a content search and does not differ from the corpus search mode in terms of its strengths and weaknesses. Accordingly, it will not receive a separate mention in the discussion of search modes below.

3.1 Workflow & Technical Implementation

KANSAS Suche 2.0 is a web-based application that is fully implemented in Java. Its workflow is illustrated in Figure 1. The basic architecture remains similar to the original *KANSAS Suche*, see Weiss et al. (2018) for comparison, but has been heavily extended in order to accommodate the additional search options offered by our system. The user can enter a search term and start a search request which is communicated from the client to the server using Remote Procedure Calls.

The front end, based on the Google Web Toolkit (GWT)[6] and GWT Material Design[7], is shown in Figure 2.[8] It allows users to choose between the three search modes: unrestricted web search, filtered web search on *alpha sites*, and corpus search

[4] `http://www.cdlponline.org`

[5] There also are literacy and basic education classes specifically designed to prepare newly immigrated people and refugees for integration courses. Our system is being designed with a focus on traditional literacy and basic education classes. While our system is not specifically targeting German as a second language learners, it is plausible to assume that our target readers include native and non-native speakers.

[6] `http://www.gwtproject.org`

[7] `https://github.com/GwtMaterialDesign`

[8] Note that the actual front end is in German. For illustration purposes, we here translated it into English.

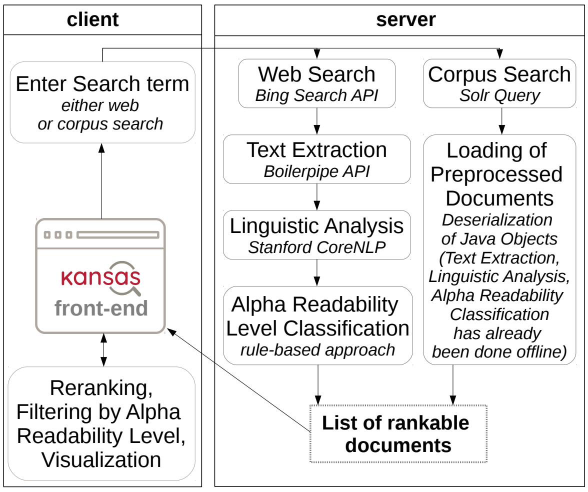


Figure 1: System workflow including web search and corpus search components

as well as the option to upload their own corpus. In the case of an unrestricted or filtered web search, the request is communicated to Microsoft Azure's BING Web Search API (version 5.0)[9] and further processed at runtime. The text content of each web page is then retrieved using the Boilerpipe Java API (Kohlschlütter et al., 2010).[10] We remove links, meta information, and embedded advertisements. The NLP analysis is then performed using the Stanford CoreNLP API (Manning et al., 2014). We identify linguistic constructions with TregEx patterns (Levy and Andrew, 2006) we defined. The linguistic annotation is also used to extract all information for the readability classification. We use the algorithm developed for *KANSAS Suche* (Weiss et al., 2018), currently the only automatic approach we are aware of for determining readability levels for low literate readers in German. The resulting list of analyzed and readability-classified documents is then returned to the client side. The user can re-rank the results based on the (de-)prioritization of linguistic constructions, filter them by Alpha Readability Level, or use the system's visualization to inspect the search results. For re-ranking we use the BM25 IR algorithm (Robertsin and Walker, 1994).

The corpus search follows a separate workflow on the server side which will be elaborated on in more detail in section 3.3 after discussing the filtered web search in section 3.2.

3.2 The Filtered Web Search

While the web provides access to a broad variety of up-to-date content, an unrestricted web search may also retrieve various types of inappropriate material. Not all search results are reading materials (sales pages, advertisement, videos, etc.), many reading materials on the web require high literacy skills, and some of the sufficiently easy reading materials contain incorrect or biased information. However, there are several web pages specialized in providing reading materials for language learners, children, or adults with low literacy skills in their native language. One option to improve web search results thus is to ensure that queries are processed so that they produce results from such web pages.

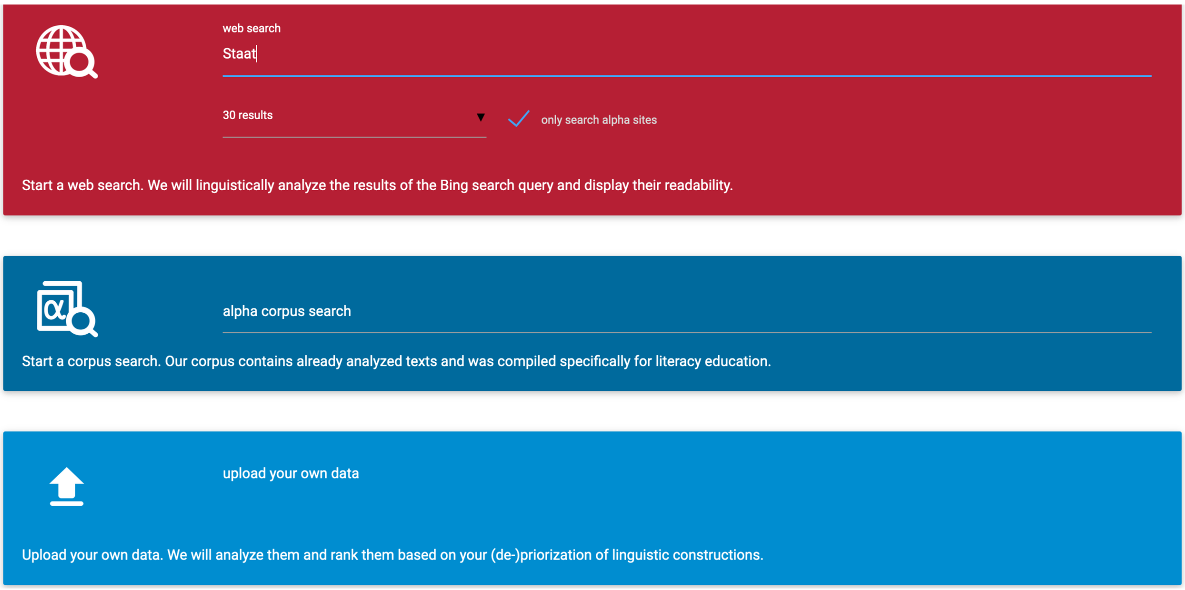
We provide the option to focus the web search on a pre-compiled list of *alpha sites*, i.e., web pages providing reading materials for readers with low literacy skills. For this, we use BING's built-in search operator `site`, which restricts the search to the specified domain. The `or` operator can be used to broaden the restriction to multiple domains. The following example illustrates this by restricting the web search for *Bundestag* ("German parliament") to the web pages of the German public international broadcaster *Deutsche Welle* (DW) and the web pages of the German Federal Agency for Civic Education *Bundeszentrale für politische Bildung* (BPB):

```
    (site:www.dw.com or
site:www.bpb.de) Bundestag
```

The `site` operator is a standard operator of most major search engines and could be directly specified using exactly this syntax by the users. In *KANSAS Suche 2.0* we integrate a special option to promote its use for a series of specific websites for three reasons. First, the `site` operator and the use of operators in search engine queries overall are relatively unknown to the majority of search engine users. Allowing users to specify a query with a `site` operator through a check box in our user interface makes this feature more accessible. Second, while specifying multiple sites is possible using the `or` operator, it becomes increasingly cumbersome the more domains are added. Having a shortcut for suitable web sites considerably increases ease of use. Third, there are a number of web pages that offer materials for low literacy classes, but many of them will not be know to the

[9] https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api

[10] https://boilerpipe-web.appspot.com



web search

Staat

30 results ▾

✓ only search alpha sites

Start a web search. We will linguistically analyze the results of the Bing search query and display their readability.

alpha corpus search

Start a corpus search. Our corpus contains already analyzed texts and was compiled specifically for literacy education.

upload your own data

Upload your own data. We will analyze them and rank them based on your (de-)priorization of linguistic constructions.

Figure 2: *KANSAS Suche 2.0* search mode options set to web query for *Staat* ("state") on 30 *alpha sites*.

user and some cannot be directly accessed by a search engine, as discussed in more detail below. Our list of *alpha sites* makes it possible to quickly access a broad selection of relevant web sites that are compatible with the functionality of *KANSAS Suche 2.0*.

To compile our list of *alpha sites*, we surveyed 75 web sites that provide reading materials for low literacy contexts. Not all of them are well-suited for the envisioned use case. We excluded web pages that offer little content (fewer than three texts), require prior registration, or predominantly offer training exercises at or below the word level rather than texts. While the latter may in principle be interesting for teachers of literacy and basic education classes, they are ill-suited for the kind of service provided by our system. The linguistic constructions that we allow the teacher or user to (de-)prioritize often target the phrase or clause level and do not make sense for individual words. However, by far the biggest drop in the number of potentially relevant web sites resulted from the fact that many web sites are designed in such a way that the materials they offer are not crawled and indexed by search engines at all. Since the material on these web sites cannot be found by search engines, it makes no sense to include them as *alpha sites* in our system.

At the end of our survey, we were left with six domains that are both relevant as accessible. This includes lexicons, news, and magazine articles in simple German (`lebenshilfe.de/de/leichte-sprache`, `hurraki.de/wiki`, `nachrichtenleicht.de`), texts written for children (`klexikon.zum.de`, `geo.de/geolino`), and texts for German as a Second Language learners (`deutsch-perfekt.com`). While this is a relatively short list compared to the number of web sites in our initial survey, these sites provide access to 34,100 materials, as identified by entering a BING search using the relevant operator specification without a specific content search term. The fact that we found so few suitable domains also showcases that the search functionality with a pre-compiled list goes beyond what could readily be achieved by a user thinking about potentially interesting sites and manually spelling out a query using the `site` operator. We are continuously working on expanding the list of *alpha sites* used by the system and welcome any information about missing options.

3.3 The Corpus Search

While there are some web sites dedicated to the distribution of reading materials for low literate readers, high-quality open source materials for literacy and basic education classes are relatively scarce. Even where materials are available, the question under which conditions teachers may alter and distribute materials often remains unclear. We want to address this issue by providing the option to specifically query for high-quality materials that have been provided as open educational resources with a corresponding license. For this, we are currently assembling a collection of such materials in collaboration with institutions creating materials for literacy and basic education.

In our system, this collection may be accessed through the same interface as the unrestricted and the filtered web search. On the server side, however, a separate pipeline is involved, as illustrated in Figure 1. Unlike the web search processing the retrieved web data on the run, the corpus search accesses already analyzed data. For this, we first perform the relevant linguistic analyses on the reading materials offline using the same NLP pipeline and readability classification as for the web search.

We use an Apache Solr index to make the corpus accessible to content queries.[11] Solr is a query platform for full-text indexing and high-performance search. It is based on the Lucene search library and can be easily integrated into Java applications. When a query request for the corpus search is sent by the user, the search results are fetched from that local Apache Solr index. In order to load the preprocessed documents into a Solr index, we transform each document into an XML file. We add a `metaPath` element which contains the name of the project responsible for the creation of the material, the author's name and the title. Additionally, we assign a `text_de` attribute to each text element, which ensures that Solr recognizes the text as German and applies the corresponding linguistic processing. The following tokenizer and filter factories, which are provided by Solr, have been set in the schema file of the index:

StandardTokenizerFactory splits the text into tokens.

LowerCaseFilterFactory converts all tokens into lowercase to allow case-insensitive matching.

StopFilterFactory removes all the words given in a predefined list of German stop words provided by the Snowball Project.[12]

GermanNormalizationFilterFactory normalizes the German special characters ä, ö, ü, and ß.

GermanLightStemFilterFactory stems the tokens using the light stemming algorithm implemented by the University of Neuchâtel.[13]

This ensures that the texts are recognized, processed, and indexed as German, which will improve the query results. Given a query, Solr returns the relevant text ids and the system can then deserialize the documents given the returned ids. Just as for the web search, the list of documents then is passed to the front-end, where the user can rerank, filter, and visualize the results.

We are still in the process of compiling the collection of high quality reading materials specifically designed for low literacy contexts. To be able to test our pipeline and evaluate the performance of the different search modes already at this stage, we use a test corpus of 10,012 texts crawled from web sites providing reading materials for low literate readers, compiled for the original *KANSAS Suche* (Weiss et al., 2018). We cleaned the corpus in a semi-automatic approach, in which we separated texts that had been extracted together and excluded non-reading materials. While we are confident that the degree of preprocessing is sufficient to demonstrate the benefits of our future corpus when compared to the web search options, it should be kept in mind that the current results are only a first approximation. The pilot corpus was only minimally cleaned so that it may still contain, for example, advertisement that would not be included in the high quality corpus being built. The pilot corpus also lacks explicit copyright information and thus is unsuitable outside of scientific analysis and demonstration purposes.

4 Comparison of Search Modes

Our system combines three search modes that have complementing strengths and weaknesses. Unrestricted web search can access a vast quantity of material, yet, most of it is not designed for low literate readers. Also, the lack of quality control may yield many unsuitable results for certain search terms, for example, those prone to elicit advertising. In contrast, restricted searches or corpus searches draw results from a considerably smaller pool of documents. Thus, although it stands to reason that the retrieved text results are of more consistent, higher quality and more likely to be at ap-

[11] `http://lucene.apache.org/solr`

[12] `http://snowball.tartarus.org`

[13] `http://members.unine.ch/jacques.savoy/clef`

propriate reading levels for our target users, there may be too few results.

To test these assumptions and see how the strengths and weaknesses play out across several queries, we compared the three search modes with regard to three criteria:

Coverage Does the search mode return enough results to satisfy a query request?

Readability Are the retrieved texts readable for low literate readers?

Suitability Are the retrieved texts suitable as teaching materials?

While the first criterion addresses a question of general interest for text retrieval systems, the other two are more specifically tailored towards the needs of our system as a retrieval system for low literacy contexts. We expect all three search modes to show satisfactory performance in general, but to exhibit the strengths and weaknesses hypothesized above.

4.1 Set-Up

For each search mode, we queried ten search terms requesting 30 results per term. The ten search terms were obtained by randomly sampling from a list of candidate terms that was compiled from the basic vocabulary list for illiteracy teaching by Bockrath and Hubertus (2014). The selection criterion for candidate terms was to identify nouns that in a wider sense relate to topics of basic education such as finance, health, politics, and society. The final list consists of the intersection of candidate terms selected by two researchers. The final ten search terms used in our evaluation are: *Alkohol* (“alcohol”), *Deutschkurs* (“German course”), *Erkältung* (“common cold”), *Heimat* (“home(land)”), *Internet* (“internet”), *Kirche* (“church”), *Liebe* (“love”), *Polizei* (“police”), *Radio* (“radio”), and *Staat* (“state”). In the following, all search terms will be referred to by their English translation.

All texts were then automatically analyzed and rated by the readability classifier used in our system. We calculated their Alpha Readability Level – both including and excluding the text length criterion of Weiss et al. (2018) based on Gausche et al. (2014); Kretschmann and Wieken (2010), since we found that many materials for low literate readers available on the web do not adhere to the text length criterion. Since texts may be relatively easily shortened by teachers before using them for literacy and basic education classes, we include both sets in our evaluation.

4.2 Coverage of Retrieved Text

Our first evaluation criterion concerns the coverage of retrieved material across search terms. While the unlimited web search (referred to as “www”) draws from a broad pool of available data, the restricted web search (“filter”) and the corpus search (“corpus”) are based on a considerably more restricted set of texts. Therefore, we first investigated to which extent the different search modes are capable of providing the requested number of results across search terms.

Overall, we obtained 817 texts for the requested 900 results. While the unrestricted web search returned the requested number of 30 results for each search term (i.e., overall 300 texts), the corpus search only retrieved 261 texts and the filtered web-search 256 texts. The latter search modes struggled to provide enough texts for the search terms *Deutschkurs* (“German course”), *Erkältung* (“common cold”). As shown in Figure 3, the corpus search returns only nine for the former and 12 results for the latter term, while the filtered search identifies seven and nine results, respectively. For the other eight search terms, all three search modes retrieve the requested 30 results.

The results indicate that with regard to plain coverage, the web search outperforms the two re-

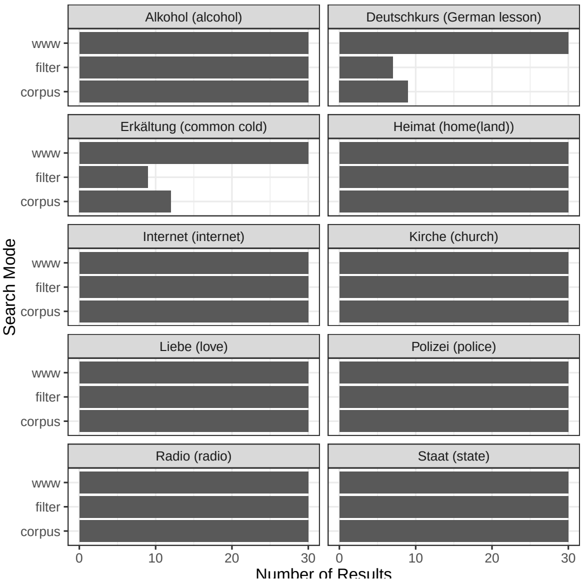


Figure 3: Results per term across search modes

stricted search modes. This is expected since neither the filtered web search nor the corpus search have access to the vast number of documents accessible to the unrestricted web search. However, they do provide the requested number of results for the majority of queries, illustrating that even the more restrictive search options may well provide sufficient coverage for many likely search terms.

4.3 Readability of Retrieved Texts

The second criterion that is essential for our comparison is the readability of the retrieved texts on a readability scale for low literacy. For this, we used the readability classifier integrated in our system to assess the Alpha Level of each text, once with and once without the text length criterion. Tables 1 and 2 show the overall representation of Alpha Readability Levels aggregated over search terms for each search mode including and ignoring text length as a rating criterion.

Alpha Level	WWW	Filter	Corpus
Alpha 3	0.00%	0.39%	4.98%
Alpha 4	19.00%	15.23%	35.25%
Alpha 5	14.33%	8.20%	14.56%
Alpha 6	10.00%	7.42%	8.43%
No Alpha	56.67%	68.75%	36.78%

Table 1: Distribution of Alpha Readability Levels *including* text length across search modes

Alpha Level	WWW	Filter	Corpus
Alpha 3	1.00%	4.30%	13.41%
Alpha 4	49.67%	53.91%	50.19%
Alpha 5	21.00%	22.66%	18.77%
Alpha 6	2.00%	10.16%	7.28%
No Alpha	26.33%	8.98%	10.34%

Table 2: Distribution of Alpha Readability Levels *ignoring* text length across search modes

As expected, the unrestricted web search elicits a high percentage of texts that are above the level of low literate readers. 56.67% of texts are rated as No Alpha and not a single text receives the rating Alpha 3. When ignoring text length, the rate of No Alpha texts drops to 26.33% but there are still only 1.00% Alpha 3 texts. It should be noted, though, that 49.67% of results are rated as Alpha 4 when ignoring text length, indicating that the unrestricted web search is not completely unsuitable for the retrieval of low literacy reading materials even though there is clear room for improvement.

The filtered web search does not seem to perform much better at first glance. On the contrary, with 68.75% it shows the overall highest rate of No Alpha labeled texts when including the text length criterion and it retrieves only 0.39% Alpha 3 texts. However, when ignoring text length, the rate of No Alpha texts drops to 8.98% – the lowest rate of No Alpha texts observed across all three search modes. It also retrieves 4.30% of Alpha 3 texts and 53.91% of Alpha 4 texts. This shows that while many of the texts found by the filtered web search seem to be too long, they are otherwise better suited for the needs of low literate readers than texts found with the unrestricted web search.

The corpus search exhibits the lowest rate of No Alpha texts (36.78%) and the highest rate of Alpha 3 and Alpha 4 texts (4.98% and 35.25%, respectively) when including the text length criterion. Without it, the rate of Alpha 3 texts even rises to 13.41%. Interestingly, though, it has a slightly higher rate of No Alpha texts than the filtered web search. That the corpus contains texts that are beyond the Alpha Levels at first may seem counterintuitive. However, the test corpus also includes texts written for language learners which may very well exceed the Alpha Levels. Considering that the majority of texts identified by this search mode are within the reach of low literate readers, this is not an issue for the test corpus. The selection of suitable materials does yield more fitting results in terms of readability.

Figure 4 shows the distribution of Alpha Readability Levels ignoring text length across search terms. It can be seen that for all search terms, Alpha Level 4 is systematically the most commonly retrieved level. A few patterns relating search terms and elicited Alpha Levels can be observed. *Deutschkurs* (“German course”), *Erkältung* (“common cold”) elicit notably fewer Alpha 4 texts, which is due to the lack of coverage in the filtered web and the corpus search. Other than that, *Polizei* (“police”) elicits by far the least No Alpha texts and among the most Alpha 3 and Alpha 4 texts, indicating that texts retrieved for this topic are overall better suited for low literacy levels. In contrast, *Radio* (“radio”) elicits most No Alpha texts and among the least Alpha 3 texts. However, it also exhibits the highest rate of Alpha 4 texts. Thus, overall it seems that the distribution

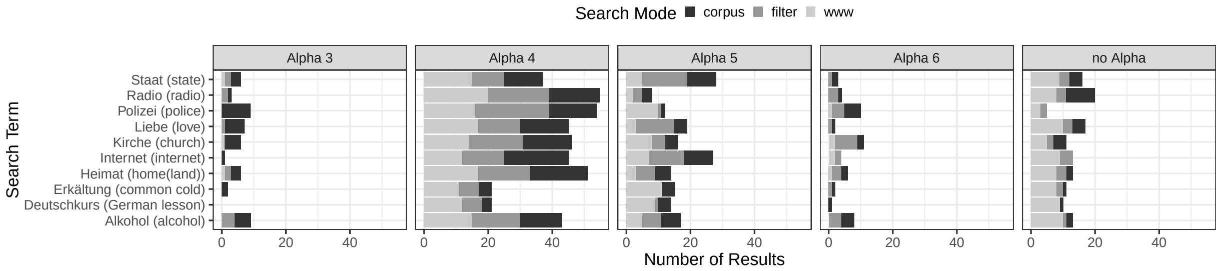


Figure 4: Distribution of query results across search terms by Alpha Level (ignoring text length)

of Alpha Readability Levels is comparable across search terms. This is in line with our expectations, given that all terms were drawn from a list of basic German vocabulary.

4.4 Suitability of Retrieved Texts

Our final criterion concerns the suitability of texts for reading purposes. As mentioned, some materials from the web are ill-suited as reading materials for literacy and basic education classes. Search results may not be reading materials but rather sales pages, advertisement, or videos. Certain search terms, such as those denoting purchasable items, such as *Radio* (“radio”), are more likely to yield such results than others. Other terms, such as those relating to politics, may be prone to elicit biased materials or texts containing misinformation. The challenge of suitability has already been recognized in previous web-search based systems, such as FERN (Walmsley, 2015) or netTrekker (Huff, 2008), where it was addressed by restricting the web search to manually verified web pages.

We investigated to which extent suitability of contents is an issue for our search modes by manually labeling materials as suitable or unsuitable on a stratified sample of the full set of queries that samples across search modes, search terms, and Alpha Readability Levels (N=451). Note that since Alpha Readability Levels are not evenly distributed across search modes, the stratified sample does not contain the same number of hits for each search term. However, each search mode is represented approximately evenly with 159 results for the corpus search, 142 for the filtered web search, and 150 for the unrestricted web search.

On this sample, we let two human annotators flag search results as unsuitable, if they were a) advertisement, b) brief captions of a video or figure, c) or a hub for other web pages on the topic. Such hub pages linking to relevant topics are not unsuitable per se, in the way advertisement or brief captions are. However, since the point of a search engine such as *KANSAS Suche 2.0* is to analyze the web resource itself rather than the pages being linked to on the page, such pages are unsuitable. Since the reliable evaluation of bias and misinformation is beyond the scope of this paper, we excluded this aspect from our evaluation. We also discarded materials as not suitable if they neither contained the search term nor a synonym to the search term. Since the information retrieval algorithm used in our corpus search is less sophisticated than the one used by BING, stemming mistakes can lead to such unrelated and thus unsuitable results. Finally, we restricted texts to 1,500 words and flagged everything beyond that as unsuitable. This is based on the practical consideration that it would take teachers too much time to review such long texts for suitability – but this rule only became relevant for six texts from the corpus, which contained full chapters from booklets on basic education matters written in simplified language.

Based on this definition of suitability that we specifically fitted for the needs of our system, our two annotators show a prevalence and bias corrected Cohen’s kappa of $\kappa = 0.765$. For the following evaluation of suitability, we only considered texts as not suitable if both annotators flagged them as such. Results that have been classified as unsuitable by only one annotator were treated as suitable materials. This resulted in overall 137 texts being flagged as not suitable, i.e., 30.38% of all search results. When splitting these results across search modes, we find that the unrestricted web search has the highest rate of unsuitable results: 52.70% of all retrieved materials were identified as not suitable. In contrast, only 8.80% of the corpus search results were labeled as unsuitable. For the filtered web search, the percentage of unsuitable materials lies between these two extremes, at 31.00%. Figure 5 shows the distribution

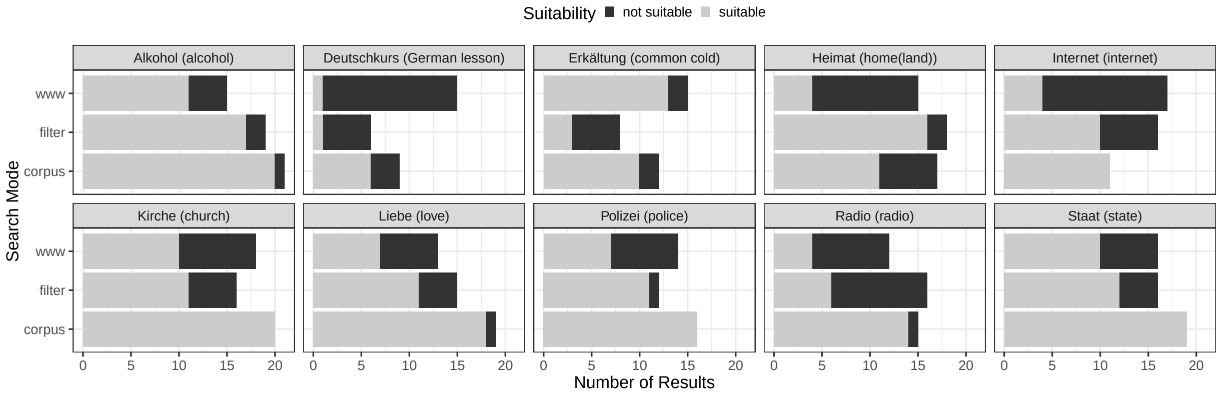


Figure 5: Suitability of sample texts (N=451) across search modes by query term

of suitable and not suitable materials across search modes split by search terms. As can be seen, some search terms elicit more unsuitable materials than others. *Deutschkurs* ("German course"), for example, contains by far more unsuitable than suitable materials for both web searches. This puts our previous findings into perspective that the unrestricted web search has higher coverage for this term than the corpus search. At least for the sample analyzed here, the corpus search retrieves considerably more suitable texts than either web search, despite its lower overall coverage.[14] The terms *Heimat* ("home(land)"), *Internet* ("internet"), and *Radio* ("radio") also seem to be particularly prone to yield unsuitable materials in an unrestricted web search.

But not all search terms elicit high numbers of unsuitable results in the unrestricted web search, see for example *Alkohol* ("alcohol"), *Erkältung* ("common cold"), and *Staat* ("state"). With for some exceptions, such as *Erkältung* ("common cold") and *Heimat* ("home(land)"), the filtered corpus search behaves similar to the unrestricted corpus search with regard to retrieving suitable materials. The corpus search clearly outperforms both web-based approaches in terms of suitability. The only term that elicits a notable quantity of unsuitable materials is *Heimat* ("home(land)"), for which the corpus includes some advertisement texts expressed in plain language. For all other search terms, corpus materials flagged as unsuitable were so based on their length.

Figure 6 displays the distribution of suitable and unsuitable materials across search modes split by readability level ignoring text length. It shows, that more than half of the Alpha Level 4 texts found in the unrestricted web search as well as approximately half of those found in the filtered web search are actually unsuitable. Similarly, a considerable number of Alpha Level 5 and nearly all Alpha Level 6 texts retrieved by the unrestricted web search in our sample are flagged as unsuitable. This puts the previous findings concerning the readability of search results into a new perspective. After excluding unsuitable results, both web searches yield considerably fewer results that are readable for low literacy levels as compared to the corpus search.

4.5 Discussion

The comparison of search modes confirmed our initial assumptions about the strengths and weaknesses of the different approaches. The unrestricted web search has the broadest plain coverage but elicits considerably more texts which require too high literacy skills or contain unsuitable materials. Although it retrieves a high proportion of Alpha 4 texts, the majority of these consist of unsuitable material. After correcting for this, it becomes apparent that users may struggle to obtain suitable reading materials at low literacy levels when solely relying on an unrestricted web search. However, depending on the search term, the rate of unsuitable materials widely differs. Thus, it stands to reason that the unrestricted web search works well for some queries while others will be less fruitful for low literacy contexts.

In these cases, the filtered web search or the corpus search can be of assistance. They both have been shown to retrieve more texts suitable for low literacy levels despite struggling with coverage

[14] This does not hold for the other term for which low coverage for all but the unrestricted web search was reported. For the search term *Erkältung* ("common cold"), the unrestricted web search finds a high number of suitable results.

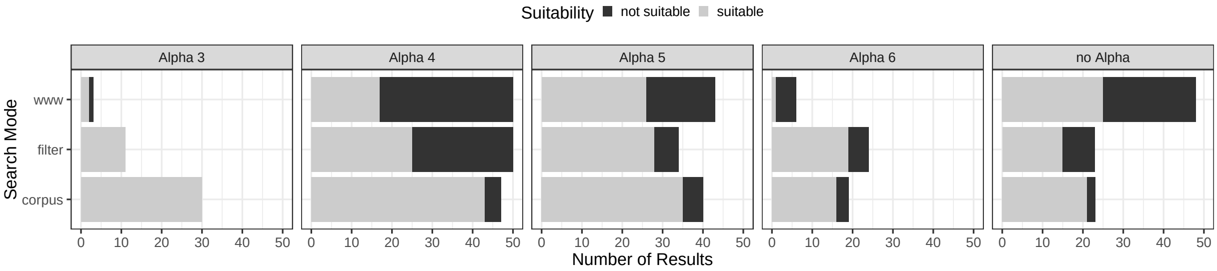


Figure 6: Suitability of sample texts (N=451) across search modes by Alpha Levels (ignoring text length)

for some search terms. Interestingly, the corpus search was shown to exceed the web search in coverage after subtracting unsuitable results for one search term. This demonstrates that raw coverage may be misleading depending on the suitability of the retrieved results. All in all, the restricted web search showed fewer advantages than the other two search modes as it suffered from both, low coverage and unsuitable materials. However, it does elicit a considerably lower ratio of unsuitable materials than the unrestricted web search, while keeping the benefit of providing up-to-date materials. Thus, we would still argue that it is a valuable contribution to the overall system.

Overall, the results show that depending on the search term and targeted readability level, it makes sense to allow users to switch between search modes so that they can identify the ideal configuration for their specific needs, as there is no single search mode that is superior across contexts.

5 Summary and Outlook

We presented our ongoing work on *KANSAS Suche 2.0*, a hybrid text retrieval system for reading materials for low literate readers of German. Unlike previous systems, our approach makes it possible to combine the strengths of unrestricted web search, broad coverage of current materials, with those of more restricted searches in curated corpora, high quality materials with clear copyright information. We demonstrated how, depending on the search term, the suitability and readability of results retrieved by an unrestricted web search can become problematic for users searching for materials at low literacy levels. Our study showed that a restricted web search and the search of materials in our corpus are valuable alternatives in these cases. Overall, there is no single best solution for all searches, so our hybrid solution allows users to choose themselves which search mode suits their needs best for a given query.

While the system itself is fully implemented, we are still compiling the corpus of reading materials for low literacy contexts and work on expanding the list of domains for our restricted web search. We are also conducting usability studies with teachers of low literacy and basic education classes and with German language teachers in training. We plan to expand the functionality of the corpus search to also support access to the corpus solely based on linguistic properties and reading level characteristics, without a content query. This will make it possible to retrieve texts richly representing particular linguistic properties or constructions that are too infrequent when having to focus on a subset of the data using the content query. We are also considering development of a second readability classifier targeting CEFR levels to accommodate the fact that German adult literacy and basic education classes are not only attended by low literate native speakers but also by German as a second language learners.

Acknowledgments

KANSAS is a research and development project funded by the Federal Ministry of Education and Research (BMBF) as part of the *AlphaDekade*[15] ("Decade of Literacy"), grant number W143500.

References

Rebekah George Benjamin. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24:63–88.

BMBF and KMK. 2016. *General Agreement on the National Decade for Literacy and Basic Skills 2016-2026. Reducing functional illiteracy and raising the level of basic skills in Germany.* Bundesministerium für Bildung und Forschung

[15] `https://www.alphadekade.de`

(BMBF), Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK). https://www.alphadekade.de/img/EN_General_Agreement_on_the_National_Decade_for_Literacy_and_Basic_Skills.pdf.

Angela Bockrath and Peter Hubertus. 2014. *1.300 wichtige Wörter. Ein Grundwortschatz*, 5th edition. Bundesverband Alphabetisierung und Grundbildung e.V., Münster, Germany.

Jonathan Brown and Maxine Eskenazi. 2004. Retrieval of authentic documents for reader-specific lexical practice. In *InSTIL/ICALL 2004 Symposium on Computer Assisted Learning, NLP and speech technologies in advanced language learning systems*, Venice, Italy. International Speech Communication Association (ISCA). http://reap.cs.cmu.edu/Papers/InSTIL04-jonbrown.pdf.

Jeanne S. Chall and Edgar Dale. 1995. *Readability revisited: the new Dale-Chall Readability Formula*. Brookline Books.

Xiaobin Chen and Detmar Meurers. 2017a. Challenging learners in their individual zone of proximal development using pedagogic developmental benchmarks of syntactic complexity. In *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa 2017*, Linköping Electronic Conference Proceedings 134, pages 8–17, Gothenburg, Sweden. ACL. http://aclweb.org/anthology/W17-0302.pdf.

Xiaobin Chen and Detmar Meurers. 2017b. Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*, 41(3):486–510.

Maria Chinkina, Madeeswaran Kannan, and Detmar Meurers. 2016. Online information retrieval for language learning. In *Proceedings of ACL-2016 System Demonstrations*, pages 7–12, Berlin, Germany. Association for Computational Linguistics. http://anthology.aclweb.org/P16-4002.

Maria Chinkina and Detmar Meurers. 2016. Linguistically-aware information retrieval: Providing input enrichment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 188–198, San Diego, CA. ACL.

K. Collins-Thompson, P. N. Bennett, R. W. White, S. de la Chica, and D. Sontag. 2011. Personalizing web search results by reading level. In *Proceedings of the Twentieth ACM International Conference on Information and Knowledge Management (CIKM 2011)*.

Scott A. Crossley, David B. Allen, and Danielle McNamara. 2011. Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, 23(1):84–101.

Frank Decroll. 1981. Funktionaler Analphabetismus – Begriff, Erscheinungsbild, psycho-soziale Folgen und Bildungsinteressen. In *Für ein Recht auf Lesen: Analphabetismus in der Bundesrepublik Deutschland*, pages 29–40.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read-it: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83.

Sukru Eraslan, Victoria Yaneva, and Yeliz Yelisada. 2017. Do web users with autism experience barriers when searching for information within web pages? In *Proceedings of the 14th Web for All Conference on The Future of Accessible Work*, pages 20–23. ACM. https://doi.org/DOI:10.1145/3058555.3058566.

B. Janghorban Esfahani, A. Faron, K. S. Roth, P. P. Grimminger, and J. C. Luers. 2016. Systematic readability analysis of medical texts on websites of german university clinics for general and abdominal surgery. *Zentralblatt fur Chirurgie*, 141(6):639–644.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 229–237, Athens, Greece. Association for Computational Linguistics. http://aclweb.org/anthology/E09-1027.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China.*

Thomas François and Cedrick Fairon. 2012. An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. https://www.aclweb.org/anthology/D12-1043.

Geert Freyhoff, Gerhard Hess, Linda Kerr, Elizabeth Menzell, Bror Tronbacke, and Kathy Van Der Veken. 1998. *Make It Simple, European Guidelines for the Production of Easy-to-Read Information for People with Learning Disability for authors, editors, information providers, translators and other interested persons*. International League of Societies for Persons with Mental Handicap European Association, Brussels.

Mark G. Friedman and Diane Nelson Bryen. 2007. Web accessibility design recommendations for people with cognitive disabilities. *Technology and Disability*, 19(4):205–2012.

S. Gausche, A. Haase, and D. Zimper. 2014. *Lesen. DVV-Rahmencurriculum*, 1 edition. Deutscher Volkshochschul-Verband e.V., Bonn.

Alex Gilmore. 2007. Authentic materials and authenticity in foreign language learning. *Language teaching*, 40(02):97–118.

Petronella Grootens-Wiegers, Martine C. De Vries, Tessa E. Vossen, and Jos M. Van den Broek. 2015. Readability and visuals in medical research information forms for children and adolescents. *Science Communication*, 37(1):89–117.

Anke Grotlüschen, Klaus Buddeberg, Gregor Dutz, Lisanne Heilmann, and Christopher 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.

Anke Grotlüschen and Wibke 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.

Anke Grotlüschen and Heike Solga. 2019. Leben mit geringer Literalität. Hauptergebnisse der LEO-studie 2018. Presentation.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, Maxine Eskenazi, Alan Juffs, and Lois Wilson. 2010. Personalization of reading passages improves vocabulary acquisition. *International Journal of Artificial Intelligence in Education*, 20:73–98.

Michael Heilman, Le Zhao, Juan Pino, and Maxine Eskenazi. 2008. Retrieval of reading materials for vocabulary and reading practice. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08*, pages 80–88, Columbus, Ohio.

Matt Huenerfauth, Lijun Feng, and Noémie Elhadad. 2009. Comparing evaluation techniques for text readability software for adults with intellectual disabilities. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*, Assets '09, pages 3–10, New York, NY, USA. ACM. http://doi.acm.org/10.1145/1639642.1639646.

Leslie Huff. 2008. Review of nettrekker di. *Language Learning & Technology*, 12(2):17–25.

Amy M. Johnson, Tricia A. Guerrero, Elizabeth L. Tighe, and Danielle S. McNamara. 2017. iSTART-ALL: Confronting adult low literacy with intelligent tutoring for reading comprehension. In *International Conference on Artificial Intelligence in Education*, pages 125–136. Springer.

J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease formula) for Navy enlisted personnel. Research Branch Report 8-75, Naval Technical Training Command, Millington, TN.

Christian Kohlschütter, Peter Frankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In *Proceedings of the Third ACM international conference on web search and data mining*, pages 441–450. ACM.

Stephen D Krashen. 1985. *The input hypothesis: Issues and implications*. Longman, New York.

R. Kretschmann and P. Wieken. 2010. *Lesen. Alpha Levels*. lea., Hamburg.

Roger Levy and Galen Andrew. 2006. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *5th International Conference on Language Resources and Evaluation*, Genoa, Italy.

Cordula Löffler and Susanne Weis. 2016. Didaktik der Alphabetisierung. In Cordula Löffler and Jens Korfkamp, editors, *Handbuch zur Alphabetisierung und Grundbildung Erwachsener*, pages 365–382. Waxmann, Münster, New York.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60. http://aclweb.org/anthology/P/P14/P14-5010.

Eleni Miltsakaki and Audrey Troutt. 2007. Read-x: Automatic evaluation of reading difficulty of web text. In *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2007*, pages 7280–7286, Quebec City, Canada. AACE. http://www.editlib.org/p/26932.

Kristina Nilsson and Lars Borin. 2002. Living off the land: The web as a source of practice texts for learners of less prevalent languages. In *Proceedings of LREC 2002, Third International Conference on Language Resources and Evaluation, Las Palmas: ELRA*, pages 411–418.

Misako Nomura, Gyda Skat Nielsen, and Bror Tronbacke. 2010. Guidelines for easy-to-read materials. revision on behalf of the ifla/library services to people with special needs section. IFLA Professional Reports 120, International Federation of Library Associations and Institutions, The Hague, IFLA Headquarters.

Ildikó Pilán, Sowmya Vajjala, and Elena Volodina. 2015. A readable read: Automatic assessment of language learning materials based on linguistic complexity. In *Proceedings of CICLING 2015- Research* *in Computing Science Journal Issue (to appear)*. https://arxiv.org/abs/1603.08868.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013a. Frequent words improve readability and short words improve understandability for people with dyslexia. In *IFIP Conference on Human-Computer Interaction*, pages 203–219, Berlin, Heidelberg. Springer. https://doi.org/DOI:10.1007/978-3-642-40498-6_15.

Luz Rello, Susana Bautista, Ricardo Baeza-Yates, Pablo Gervás, Raquel Hervás, and Horacio Saggion. 2013b. One half or 50%? an eye-tracking study of number representation readability. In *IFIP Conference on Human-Computer Interaction*, pages 229–245, Berlin, Heidelberg. Springer. https://doi.org/DOI:10.1007/978-3-642-40498-6_17.

Stephen Robertsin and Steve Walker. 1994. Some simple effective approximations to the 2-poissin model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 131–241.

Hannes Schröter and Irit Bar-Kochva. 2019. Keyword: Reading literacy. Reading competencies in Germany and underlying cognitive skills. *Zeitschrift für Erziehungswissenschaft*, 22(1):17–49.

Sanja Štajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov. 2012. What can readability measures really tell us about text complexity? In *In Proceedings of the First Workshop on Natural Language Processing for Improving Textual Accessibility*. European Language Resources Association (ELRA).

Merrill Swain. 1985. Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In Susan M. Gass and Carolyn G. Madden, editors, *Input in second language acquisition*, pages 235–253. Newbury House, Rowley, MA.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 163–173, Montréal, Canada. ACL. http://aclweb.org/anthology/W12-2019.pdf.

Sowmya Vajjala and Detmar Meurers. 2013. On the applicability of readability models to web texts. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 59–68.

Sowmya Vajjala and Detmar Meurers. 2014. Readability assessment for text simplification: From analyzing documents to identifying sentential simplifications. *International Journal of Applied Linguistics, Special Issue on Current Research in Readability and Text Simplification*, 165(2):142–222.

Michael Walmsley. 2015. *Learner Modelling for Individualised Reading in a Second Language*. Ph.D. thesis, The University of Waikato. http://hdl.handle.net/10289/10559.

Zarah Weiss, Sabrina Dittrich, and Detmar 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.

Zarah Weiss and Theresa Geppert. 2018. Textlesbarkeit für Alpha-Levels. Annotationsrichtlinien für Lesetexte. Version 1.1. http://www.sfs.uni-tuebingen.de/~zweiss/rsrc/textlesbarkeit-fur-alpha.pdf.

Zarah Weiss and Detmar Meurers. 2018. Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, Santa Fe, New Mexico, USA. https://www.aclweb.org/anthology/C18-1026.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Victoria Yaneva. 2015. Easy-read documents as a gold standard for evaluation of text simplification output. In *Proceedings of the Student Research Workshop associated with RANLP 2015*, pages 30–36, Hissar, Bulgaria.

Victoria Yaneva, Irina Temnikova, and Ruslan Mitkov. 2016a. Accessible texts for autism: An eye-tracking study. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, pages 49–57.

Victoria Yaneva, Irina P. Temnikova, and Ruslan Mitkov. 2016b. Evaluating the readability of text simplification output for readers with cognitive disabilities. In *Proceedings of the 10h International Conference on Language Resources and Evaluation (LREC)*, pages 293–299.