A Linguistically-Informed Search Engine to Identifiy Reading Material for Functional Illiteracy Classes

Zarah Weiss **Sabrina Dittrich** **Detmar Meurers**
Department of Linguistics, ICALL-Research.de Group
LEAD Graduate School & Research Network
University of Tübingen
{zweiss,dm}@sfs.uni-tuebingen.de,
sabrina.dittrich@uni-tuebingen.de

Abstract

We present KANSAS, a search engine designed to retrieve reading materials for functional illiterates and learners of German as a Second Language. The system allows teachers to refine their searches for teaching material by selecting appropriate readability levels and (de)prioritizing linguistic constructions. In addition to this linguistically-informed query result ranking, the system provides visual input enhancement for the selected linguistic constructions.

Our system combines state-of-the-art Natural Language Processing (NLP) with light-weight algorithms for the identification of relevant linguistic constructions. We have evaluated the system in two pilot studies in terms of the identification of linguistic constructions and the identification of readability levels. Both pilots achieved highly promising results and are being followed by full-fledged performance studies and usability tests.

1 Introduction

We present KANSAS, a linguistically-informed search engine designed to support teachers for adult literacy and German as a Second Language (GSL) classes in their search for appropriate reading materials.[1] Functional illiteracy describes the inability to read or write short coherent texts. This includes the inability to comprehend everyday reading materials such as information brochures or operating instructions. It is a pressing issue for modern society; approximately 7.5 million people in Germany are functional illiterates, which corresponds to 14.5% of the working-age population (18-64 years) (Riekmann and Grotlüschen, 2011). For teachers of adult literacy classes, it is particularly difficult to find reading material that is appropriate for their students. While the need for authentic reading material with particular linguistic characteristics has also been pointed out for foreign language teaching (Chinkina et al., 2016), the issue in the functional illiteracy context is even more pressing given that adult literacy classrooms are highly culturally and linguistically diverse. Learners have heterogeneous biographical and educational backgrounds, they may or may not be native speakers of German, and their low literacy skills may or may not be associated with a cognitive disability, which is commonly considered to include, among others, populations with Autism Spectrum Disorders (ASD), dyslexia, intellectual disorders, traumatic brain injuries, aphasia, dementia, Alzheimer's disease, and Attention Deficit (Hyperactivity) Disorder (Friedman and Bryen, 2007; Huenerfauth et al., 2009). This substantial diversity has to be considered when selecting teaching materials, also making the use of textbooks particularly questionable. In practice, adult literacy teachers depend on identifying appropriate materials for their classes online using standard content search engines like *Google* or *Bing*. However, identifying adequate reading material for readers with lower reading skills is a challenging task: Huenerfauth et al. (2009) and Feng (2009) point out that many texts that are accessible at low literacy levels actually target children and their content may thus be ill-suited for adult readers; texts of interest to adult readers often require higher levels of literacy. Vajjala and Meurers (2013) show that the reading level of web query results obtained using *Bing* is variable, but on average quite high. Web content specifically designed for readers with low reading skills is not necessarily suited for all learners either, due to the diversity of conditions that result in low literacy

[1] https://www.kansas-suche.de/

Zarah Weiss, Sabrina Dittrich and Detmar Meurers 2018. A linguistically-informed search engine to identifiy reading material for functional illiteracy classes. *Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning at SLTC 2018 (NLP4CALL 2018).* Linköping Electronic Conference Proceedings 152: 79–90.

skills (Yaneva, 2015). Our system is designed to support teachers in this challenging task of identifying appropriate material by combining content queries with the flexible (de)prioritization of relevant linguistic constructions and filtering results by readability levels.

The system design is based on insights from Second Language Acquisition (SLA) research. Similar to SLA, the acquisition of reading and writing skills, even in the L1, does not happen implicitly through exposure but through explicit instruction. Thus, insights from SLA research are highly relevant for the context of literacy training. The importance of *input* for successful language acquisition is well-established in SLA research (Krashen, 1977; Swain, 1985). According to Krashen's *Input Hypothesis* (Krashen, 1977), learning is facilitated by exposure to input that is slightly more advanced than a learner's current state of language competence (*i+1*). We promote the identification of appropriate texts by offering a readability level filter that is designed to specifically target the reading competence of functional illiterates. Another insight from SLA research that we included in the design of our system is that the salience of linguistic constructions and the recognition of these constructions by the learner is a crucial component of language learning, as established by Schmidt's *Noticing Hypothesis* (Schmidt, 1990). One prominent approach to promote salience of linguistic constructions is (visual) *input enhancement* (Smith, 1993) in terms of, e.g., colors, font changes, or spacing. KANSAS integrates these two aspects by i) giving users the option to promote search results that contain relevant linguistic constructions and by ii) visually enhancing these constructions in the reading text. By taking the perspective of SLA research into consideration, we also approach a broader group of learners, including GSL. This matches the reality of most German literacy classes, which are not only attended by native speakers with reading deficiencies but also by some non-native speakers. Also, while KANSAS is designed for educational purposes and focuses on the functional illiterate reading population, it can also facilitate the identification of well-suited reading materials in ordinary web searches conducted by users with low literacy skills, who face the same issues as literacy teachers when it comes to the identification of accessible reading materials (Eraslan et al., 2017; McCarthy and Swierenga, 2010).

The article is structured as follows: First, we give some background on related work. In Section 3, we then describe our system's technical implementation and general workflow. We put a special focus on its two main components: the algorithm for the identification of relevant linguistic constructions and the readability assessment algorithm. We then present the preliminary evaluation of these two algorithms from two pilot studies, which are currently being extended by follow up studies. We conclude with an outlook on future steps.

2 Background

In addition to other information retrieval systems that have been designed for the purpose of language acquisition, our work heavily draws on previous work on readability assessment in the context of SLA research, research on the accessibility of reading materials for users with cognitive disabilities, and specifically on German illiteracy research.

2.1 Related Systems

The idea of retrieving and making use of authentic web texts for language learning purposes has been investigated in several research approaches.

The ICALL systems *VIEW* and *WERTi* provide input enhancement techniques for websites (Meurers et al., 2010). They support visually enhancing selected linguistic constructions in order to make them more salient to the learner. Furthermore, they automatically generate fill-in-the-gap exercises for these constructions and embed them into the websites in real-time.

Another productive line of research investigates the design of search engines for language learners. The *REAP* tutoring system (Brown and Eskenazi, 2004) helps selecting appropriate reading material from a digital library data base by matching texts against a student model focusing on vocabulary acquisition. It has also been ported to Portuguese (Marujo et al., 2009). Ott and Meurers (2011) developed *LAWSE*, a search engine prototype that takes reading difficulty measures into account. A similar system is READ-X (Miltsakaki and Troutt, 2007), a search engine that analyzes text readability by making use of traditional readability formula.

Finally, the *FLAIR* system (Form-Focused Linguistically Aware Information Retrieval) by Chinkina et al. (2016) emphasizes the importance of including grammar knowledge into such information retrieval systems. *FLAIR* integrates grammatical patterns specified in an official English L2 class curriculum into a content-based search engine. The system allows users to rerank search results by assigning weights to linguistic constructions. Furthermore, it visually enhances these constructions in a simple reading view and allows to filter texts for readability based on a readability formula. KANSAS adapts *FLAIR* to German and focuses primarily on the special needs of functional literacy training.

2.2 Readability Assessment

Readability assessment is the task of matching texts to readers of a certain population based on the (linguistic) complexity of the text. The earliest approach is the use of simple readability formulas such as the Flesch-Kincaid formula (Kincaid et al., 1975) or the Dale-Chall readability formula (Chall and Dale, 1995); see DuBay (2006) for an overview. These formulas are still widely used in non-linguistic studies (Esfahani et al., 2016; Grootens-Wiegers et al., 2015) and in information retrieval systems (cf. Section 2.1). However, readability formulas are known to be highly limited and potentially unreliable as they only capture superficial text properties such as sentence and word length (Feng et al., 2009; Benjamin, 2012). Research on readability assessment thus has shifted towards broader linguistic modeling of syntactic, lexical, and discourse complexity based on elaborate Natural Language Processing (NLP) pipelines and successfully adopted features from SLA research (Feng et al., 2010; Vajjala and Meurers, 2012). Measures of discourse and textual cohesion were also shown to be highly relevant for readability assessment (Crossley et al., 2008, 2011; Feng et al., 2009), as well as psycho-linguistic measures of language use (Chen and Meurers, 2017; Weiss and Meurers, 2018). While most work on readability assessment was conducted for English, the findings have also been corroborated for other languages such as French (François and Fairon, 2012), Italian (Dell'Orletta et al., 2011), and German (Vor der Brück et al., 2008; Hancke et al., 2012; Weiss and Meurers, 2018).

These data-driven machine learning approaches to readability modeling are not feasible for these populations due to a lack of (labeled) training data (Yaneva et al., 2016). Although there are corpus-based approaches to comparative readability assessment for low literacy readers (cf., e.g., Feng et al., 2009; Yaneva et al., 2016), eye-tracking studies are more common in research on readability assessment for these groups: Rello et al. investigate the effect of noun frequency and noun length (Rello et al., 2013a) and the effect of number representations (Rello et al., 2013b) on the readability and comprehensibility of texts for Spanish L1 readers with dyslexia. Eraslan et al. (2017) investigate general information extraction strategies of users with high functioning autism on web pages using eye-tracking and Yaneva et al. (2015) employ eye-tracking to study attention patterns of readers with ASD in contextualized documents containing images as well as text material. They derive recommendations from their findings to improve text accessibility for readers with low literacy skills. Among other things, they recommend the use of plain English matching Easy-to-Read requirements as suitable in their complexity for readers with ASD. With this, they link eye-tracking research to another increasingly popular approach for the evaluation of reading materials for populations with cognitive disabilities: the adherence to guidelines for the production of Easy-to-Read materials. Easy-to-Read materials are specifically designed to enhance the accessibility of texts for readers with cognitive disabilities; examples are the guidelines by Nomura et al. (2010) and Freyhoff et al. (1998). These guidelines comment on text layout as well as on language complexity. Yaneva (2015) operationalizes some of the language-focused recommendations in Freyhoff et al. (1998)'s Easy-to-Read guidelines in terms of automatically accessible linguistic features. She uses the resulting algorithm to evaluate web material marked as Easy-to-Read document in terms of their compliance to these guidelines and their similarity to material specifically designed for two target populations of Easy-to-Read language: readers with ASD and readers with mild ID. Yaneva et al. (2016) use this algorithm to evaluate reading materials for readers with cognitive disabilities in terms of their compliance to Easy-to-Read standards.

2.3 Functional Illiteracy

Two major studies have addressed the issue of functional illiteracy in Germany: The *lea. - Literalitätsentwicklung von Arbeitskräften* study ("literacy development for workers") and the *leo. - Level-One* study.[2] They defined degrees of (functional) illiteracy and severely low reading and writing abilities. They define functional illiteracy as reading and writing skills at which individual sentences may be written or read, but not coherent texts even if they are short. Severely low reading and writing abilities are above the level of functional illiteracy, but at this level literacy competence is still highly limited and does not exceed short or intermediate texts. In the course of these studies, the so called *Alpha Levels* were developed to systematically address degrees of limited literacy in the German population (Riekmann and Grotlüschen, 2011). Alpha levels range from Alpha 1 to Alpha 6. Reading and writing skills at Alpha Levels 1 to 3 constitute functional illiteracy, while Alpha Levels 4 to 6 describe varying degrees of low literacy. Table 1 displays the reading skill dimension of these levels.

We used these descriptions of reading and writing competencies across Alpha Levels to derive corresponding criteria reading materials have to adhere to in order to be suitable for the respective Alpha Levels. We excluded Alpha Levels 1 and 2, because these only apply to the character and word level and are thus not applicable to queries for texts. We henceforth refer to these reading levels as *Alpha readability levels* (Alpha 3 to 6 and above Alpha). We elaborate on our approach in Section 3.3.

3 System Description

KANSAS focuses on the reranking of content queries based on the prioritization of specific grammatical constructions. With this, we follow the approach outlined by Chinkina et al. (2016). For this, we ported some linguistic constructions from *FLAIR* to German and implemented new constructions that are relevant to the contexts of German illiteracy and L2 reading acquisition. Furthermore, we introduced the de-prioritization of grammatical constructions into our system to accommodate for the special needs of adult literacy teaching contexts. As previous systems, we also provide reading level based filtering of texts. However, unlike previous information retrieval systems, we go beyond simple readability formulas and employ a more linguistically-informed approach to readability assessment.

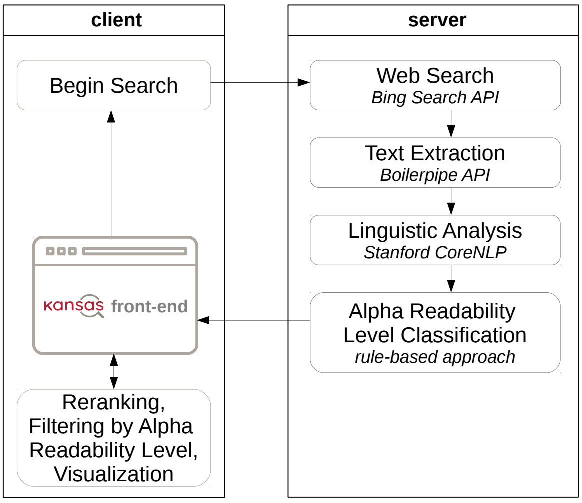


Figure 1: Overview of the KANSAS's workflow.

3.1 Technical Implementation

KANSAS is a web-based application developed in Java using the Google Web Toolkit (GWT). The technical architecture including web search, crawling, parsing, and ranking is based on *FLAIR* (Chinkina et al., 2016): Remote Procedure Calls (RPC) are used for client server communication. The BING Web Search API version 5.0[3] is employed for the web search and the Boilerpipe Java API[4] for text extraction. The linguistic preprocessing is performed using Stanford CoreNLP.[5] The BM25 IR algorithm (Robertson and Walker, 1994) is used to combine the weights for content fit and linguistic constructions. For the front-end design, we use GWT Material Design[6].

3.2 Workflow

Figure 1 illustrates our system architecture and workflow. While the system's basic architecture strongly resembles the *FLAIR* pipeline described in Chinkina et al. (2016), we did not merely reimplement *FLAIR*. We systematically redesigned the components web search, text extraction, linguistic analysis, and ranking to German, and ex-

[2] http://blogs.epb.uni-hamburg.de/lea/, http://blogs.epb.uni-hamburg.de/leo/.

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

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

[5] https://stanfordnlp.github.io/CoreNLP/

[6] https://github.com/GwtMaterialDesign

Level	Reading skills
Alpha 1	pre-literal reading (character level)
Alpha 2	constructs meaning at word level
Alpha 3	constructs meaning at sentence level
Alpha 4	constructs meaning at test level and knows high-frequent words
Alpha 5/6	increasingly literate at intermediate text length

Table 1: Definition of Alpha Levels (cf. Riekmann and Grotlüschen, 2011, p. 28, Table 1).

tended them to the special needs context of adult literacy teaching. Furthermore, we developed a readability filter performing a refined and empirically grounded classification of texts into Alpha readability levels.

Web search. The workflow starts with the client sending a search query to the server. On server side, the BING Web Search API is prompted to query for relevant search results. While *FLAIR* filters these results by discarding all texts containing less than 100 words, we set the lower word limit to 10 words and additionally discard all texts with more than 400 words as these are necessarily unsuited for adult literacy classes.

Text extraction. To remove boilerplate and template strings that do not belong to the websites' main textual content, we make use of the *ArticleExtractor* included in the Boilerpipe Java API. We chose this extractor, which has been trained on news articles, after piloting the performance of all available filters.

Linguistic analysis/preprocessing. We use the Stanford CoreNLP API to extract linguistic annotations from the resulting plain texts. We use the German shift-reduce model for parsing.

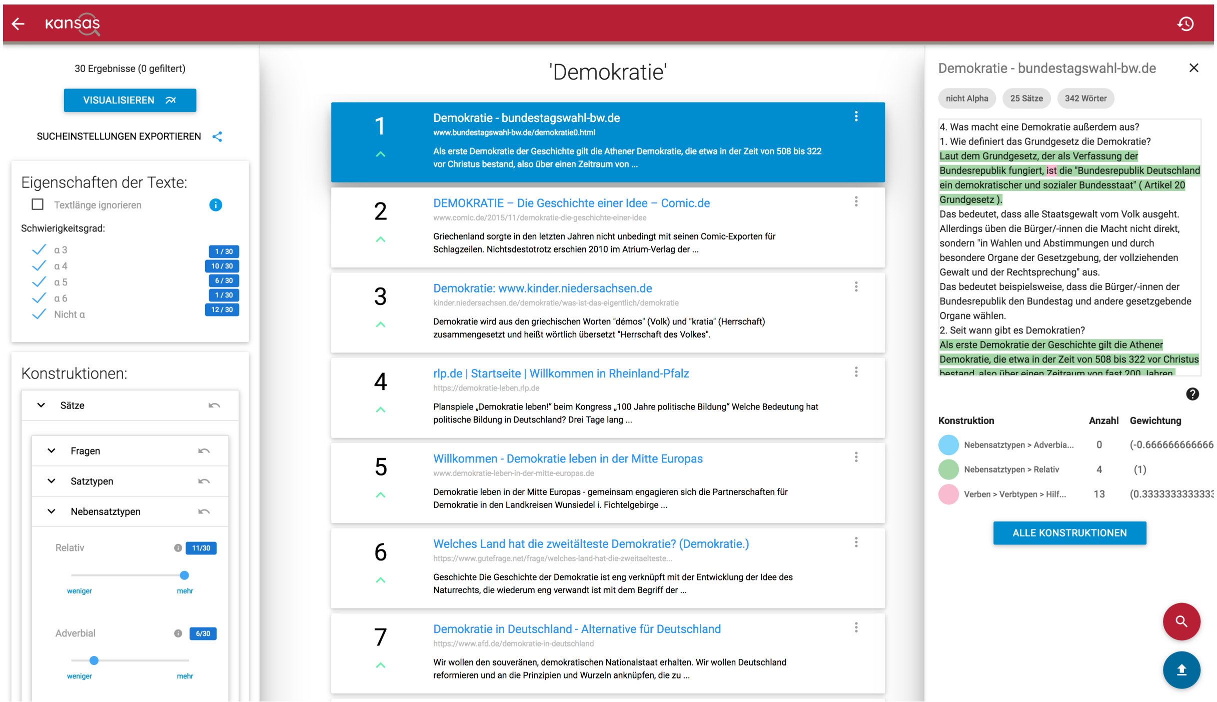
Alpha level classification. Based on the linguistic analysis, we compute a set of features to determine a text's Alpha readability level. We assign these levels to texts following a rule-based approach, which is outlined in more detail in Section 3.3 and evaluated in Section 4.2.

Ranking, filtering, and visualization. On the client side, the user is asked to wait until the analysis is completed. Afterwards, the user can inspect the linguistically analyzed query results. Figure 2 shows how the results are displayed to the user: The settings panel on the left contains range sliders that allow the user to set priority weights to a broad range of linguistic constructions. Setting a construction's weight to a negative value penalizes texts containing the construction, while positive values cause higher ranks. Each time a slider is changed, the results are reranked accordingly and the construction gets highlighted in the text preview window on the right. This may either be used for verification of the automatic analysis or as visual enhancement for teaching purposes. The performance of this feature is evaluated in Section 4.1. Additionally, user may filter query results for certain Alpha readability levels. We also re-implemented *FLAIR*'s visualization perspective which allows to inspect the occurrences of constructions across texts.

3.3 Main Algorithms

KANSAS is based on two main algorithms: The first algorithm concerns the extraction of linguistic constructions from a textual document. This algorithm is relevant for two important functionalities: First, users are given the possibility to rank search results by prioritizing and de-prioritizing certain linguistic constructions. Second, the constructions are visually enhanced within the text preview (cf. Figure 2). The second algorithm classifies texts into Alpha readability levels.

The algorithm for the detection of the constructions is based on our NLP preprocessing pipeline. In total, 85 construction types are annotated on sentence-, phrase-, or token-level based on part-of-speech (POS) annotations and constituency trees. On the sentence-level, we extract sentence types (e.g., simple or complex sentences) and question types (e.g., wh-questions). On the phrase-level, subordinate clause types (e.g., relative clauses) are extracted. On the word-level, we annotate properties of verbs, adjectives, nouns, negations, determiners, pronouns and prepositions. We use Tregex to identify patterns in parse trees based on regular expressions (Levy and Andrew, 2006). While *FLAIR*, too, makes use of Tregex patterns, we newly implemented all patterns to fit the German syntax and POS tags. We excluded constructions that are not relevant for German, such as long and short form adjective comparative con-



☐ Textlänge ignorieren

Schwierigkeitsgrad:

✓ α 3 1 / 30
✓ α 4 10 / 30
✓ α 5 6 / 30
✓ α 6 1 / 30
✓ Nicht α 12 / 30

Relativ 11/30

Adverbial 6/30

1 Demokratie - bundestagswahl-bw.de
www.bundestagswahl-bw.de/demokratie0.html
Als erste Demokratie der Geschichte gilt die Athener Demokratie, die etwa in der Zeit von 508 bis 322 vor Christus bestand, also über einen Zeitraum von ...

2 DEMOKRATIE – Die Geschichte einer Idee – Comic.de
www.comic.de/2015/11/demokratie-die-geschichte-einer-idee
Griechenland sorgte in den letzten Jahren nicht unbedingt mit seinen Comic-Exporten für Schlagzeilen. Nichtsdestotrotz erschien 2010 im Atrium-Verlag der ...

3 Demokratie: www.kinder.niedersachsen.de
kinder.niedersachsen.de/demokratie/was-ist-das-eigentlich/demokratie
Demokratie wird aus den griechischen Worten "démos" (Volk) und "kratia" (Herrschaft) zusammengesetzt und heißt wörtlich übersetzt "Herrschaft des Volkes".

4 rlp.de | Startseite | Willkommen in Rheinland-Pfalz
https://demokratie-leben.rlp.de
Planspiele „Demokratie leben!“ beim Kongress „100 Jahre politische Bildung“ Welche Bedeutung hat politische Bildung in Deutschland? Drei Tage lang ...

5 Willkommen - Demokratie leben in der Mitte Europas
www.demokratie-leben-in-der-mitte-europas.de
Demokratie leben in der Mitte Europas - gemeinsam engagieren sich die Partnerschaften für Demokratie in den Landkreisen Wunsiedel i. Fichtelgebirge ...

6 Welches Land hat die zweitälteste Demokratie? (Demokratie.)
https://www.gutefrage.net/frage/welches-land-hat-die-zweitaelteste...
Geschichte Die Geschichte der Demokratie ist eng verknüpft mit der Entwicklung der Idee des Naturrechts, die wiederum eng verwandt ist mit dem Begriff der ...

7 Demokratie in Deutschland - Alternative für Deutschland
https://www.afd.de/demokratie-in-deutschland
Wir wollen den souveränen, demokratischen Nationalstaat erhalten. Wir wollen Deutschland reformieren und an die Prinzipien und Wurzeln anknüpfen, die zu ...

Demokratie - bundestagswahl-bw.de

nicht Alpha 25 Sätze 342 Wörter

4. Was macht eine Demokratie außerdem aus?
1. Wie definiert das Grundgesetz die Demokratie?
Laut dem Grundgesetz, der als Verfassung der Bundesrepublik fungiert, ist die "Bundesrepublik Deutschland ein demokratischer und sozialer Bundesstaat" (Artikel 20 Grundgesetz).
Das bedeutet, dass alle Staatsgewalt vom Volk ausgeht.
Allerdings üben die Bürger/-innen die Macht nicht direkt, sondern "in Wahlen und Abstimmungen und durch besondere Organe der Gesetzgebung, der vollziehenden Gewalt und der Rechtsprechung" aus.
Das bedeutet beispielsweise, dass die Bürger/-innen der Bundesrepublik den Bundestag und andere gesetzgebende Organe wählen.
2. Seit wann gibt es Demokratien?
Als erste Demokratie der Geschichte gilt die Athener Demokratie, die etwa in der Zeit von 508 bis 322 vor Christus bestand, also über einen Zeitraum von fast 200 Jahren

Konstruktion	Anzahl	Gewichtung
● Nebensatztypen > Adverbia...	0	(-0.666666666666
● Nebensatztypen > Relativ	4	(1)
● Verben > Verbtypen > Hilf...	13	(0.333333333333

Figure 2: KANSAS's interface: This view displays the search results for the query *Demokratie* ("democracy"). On the settings panel on the left, the user can assign weights to linguistic constructions and filter for Alpha readability levels. The preview panel on the right highlights selected constructions.

structions. We also implemented new constructions that are specifically relevant for the contexts of German and adult literacy classes, such as various constructions used for the elaboration of the German nominal domain and verb position features. The performance of this algorithm is evaluated in sections 4.1.

The second crucial algorithm employed in KANSAS is a sophisticated readability filter for Alpha readability levels. In order to find texts that match the reading skills of the intended target group, we developed a theoretically grounded algorithm to identify readability levels for functional illiterates. We based this rule-based algorithm on the operationalization of criteria for the identification of functional illiteracy levels (Alpha 3 to Alpha 6) (cf. Section 2.3). We used the detailed ability-based descriptions provided by Gausche et al. (2014) and Kretschmann and Wieken (2010) to derive robust operationalizations of each Alpha Level in terms of concrete text characteristics along the dimensions of text length, sentence length, sentence structure, tense patterns, and word length and extract all linguistic features relevant for this assessment from our NLP preprocessing pipeline.[7] We preferred this approach over one adopting guidelines for Easy-to-Read materials as done in previous work (cf. Section 2.2). Following the ability-based descriptions of degrees of functional illiteracy allows us to differentiate reading levels within the reach of readers with low literacy skills. Furthermore, unlike text production guidelines, German *Alpha Levels* specify concrete thresholds for most of their linguistic characteristics, which allows us to evaluate materials without using reference corpora containing reading materials that were verified to be suited for readers with low literacy skills. This is crucial for our approach given that such corpora are not freely available for German.

4 System Evaluation

We have evaluated both of KANSAS's core algorithms in two pilot studies. First, we tested the performance of our linguistic construction identification algorithm for a subset of five linguistic constructions. Second, we evaluated the performance of our readability assessment algorithm by comparing it to the performance of a human expert annotator.

[7] The complete algorithm may be found in the Appendix in Figure 3.

4.1 Identification of Linguistic Constructions

We analyzed five target constructions from our list of overall 85 linguistic constructions. We chose four constructions that are extracted using Tregex patterns, because these are more elaborate and thus more prone to errors. We also chose one construction that is solely based on Stanford CoreNLP POS tags to compare its performance to the other constructions. Furthermore, we only chose constructions that are particularly relevant for adult literacy classes. This resulted in the following target constructions:

Complex sentences are sentences that contain more than one clause, e.g., *Ich spiele und du liest* ("I am playing and you are reading").

***Haben* perfect** is the simple perfect formed with *haben* ("to have"), e.g., *Ich habe geschlafen* ("I have slept").

Participle verbs are verbs in the non-finite form that is used to form periphrastic tenses such as simple perfect and past perfect.

Adjectival attributes are adjectives that are attributes to noun phrases, e.g., *der* ***grüne*** *Ball* ("the green ball").

LSB + RSB clauses are clauses that contain at least two verb components which are separated by an arbitrary amount of language material in the center of the clause, e.g., *Sie* ***hat*** *in der Mensa* ***gegessen***. ("She ate in the canteen").[8]

To evaluate how robustly the algorithm identifies these constructions, we analyzed five to ten articles for each target construction. We performed queries with our system for several search terms and selected the highest ranking of 40 documents after re-ranking the query results by prioritizing the respective target construction.[9] We collected articles until we observed a sufficient amount of instances for each target construction (15 to 59). Table 2 reports precision, recall, and f-measure for each target construction as well as the amount of observed constructions on which the results are based. On average we observe a satisfactory performance across all target constructs. However, the low recall we observe for adjectival noun attributes ($rec. = .593$) indicates that our algorithm may yet be improved. A qualitative analysis of the false negative instances showed that in coordinated adjectival noun attributes the second adjectival attribute is often but not always missed by the algorithm. We are currently investigating the cause for this. However, this issue is less pressing for the system's overall performance, since high precision is more important for the prioritization and visual enhancement of target constructions.

Construction	N	Prec	Rec	F_1
Complex sentences	43	.788	.953	.863
haben-perfect	15	1.00	.867	.929
Participle verbs	42	.929	.929	.929
Adjectival attributes	59	.946	.593	.729
LSB + RSB	31	.893	.806	.847
Mean score	38	.911	.830	.859

Table 2: Performance of identification of linguistic constructions.

Overall, these preliminary findings are encouraging and give us crucial insights into which aspects of our algorithm require more performance tuning. We are continuing to evaluate all constructions identified by KANSAS and to further improve on our construction identification algorithm.

4.2 Identification of Readability Levels

We conducted a preliminary evaluation of our readability level filter by matching its ratings against human expert judgments in terms of interrater reliability. For this, we crawled $N = 68$ texts from websites that offer reading materials for functional illiterates and German L2 learners. We let a human annotate these texts, who was considered an expert because she had extensively studied the ability-based descriptions of functional illiteracy levels by Gausche et al. (2014) and Kretschmann and Wieken (2010) as well as the example material provided by them in the months prior to the annotation procedure. The human annotations were based on annotation guidelines that we derived from the same ability-based Alpha Level descriptions we used for the design of our rule-based algorithm (Weiss and Geppert, 2018).

[8] We refer to this type of clause as LSB + RSB clause as a shorthand for left sentence bracket + right sentence bracket clauses, which are names for the respective positions of the verb components in the *Topological Field Model* (Wöllstein, 2014).

[9] We used the following query terms: *Demokratie* ("democracy"), *Bundestag* (the German federal parliament), *Chancengleichheit* ("equal opportunity"), and *Bildungsmassnahme* ("educational measures").

We then automatically rated the same texts with our Alpha readability classifier and calculated the inter-rater reliability (IRR) of the ratings. This procedure allowed us to obtain a preliminary evaluation of the performance of our algorithm despite the lack of a suited Gold Standard.

Before we calculated the IRR, we tested for prevalence using the Stuart Maxwell test for marginal homogeneity but did not find any significant prevalence. We also tested for rater bias by calculating the coefficient of systematic bias between two raters but did not find any significant bias. Accordingly, we calculated Cohen's κ (Cohen, 1960) and observed substantial agreement between the human expert and our algorithm (κ = .63). We additionally calculated weighted κ_w (Cohen, 1968) in order to account for the ordinal structure in our data. Following Hallgren (2012) we chose quadratic weights to differentiate between degrees of disagreement between two raters. We observe near perfect agreement for quadratic weighted κ (κ_w = .90). All analyses were conducted using used the R package IRR (v. 0.84).[10]

While the described procedure is only an initial pilot study, which is limited in terms of its validity due to the lack of a second annotator, it already shows highly promising results. We are now addressing the limitations of the pilot by evaluating the robustness of the readability algorithm as well as of our human rater guidelines in a more elaborate study with 300 additional texts rated by two human annotators.

5 Conclusion & Outlook

KANSAS is the first web search engine designed to identify texts for German functional illiterates or German as a Second Language. The system supports the flexible (de)prioritization and visual enhancement of 85 linguistic constructions that are important for German adult literacy teaching and GSL learning contexts. Our theoretically grounded readability algorithm is specifically calibrated towards the needs of functional illiterates. It thus addresses the issue that most reading materials that may be found on the Internet are ill-suited for the special reading needs of functional illiterates.

We presented KANSAS's main features and evaluated its key algorithms in two pilot studies. Our exemplary analysis of the performance of the identification of linguistic constructions shows a promising overall performance with high f-scores across four out of five constructions ranging from 0.85 to 0.93. The rule-based algorithm which rates the readability of texts was compared with the performance of a human expert annotator. We observed high agreement results with a Cohen's κ value of 0.63 and weighted κ_w of 0.9. We tuned our readability algorithm specifically towards the target group of German functionally illiterates by basing it on the German official criteria for the identification of functional illiteracy levels.

Our pilot studies successfully demonstrate the robustness of our algorithms in real-life applications. The web system is platform-independent and freely available online. While some of the functionality is also featured in previous work on the *FLAIR* system for English, we also provide novel features such as a sophisticated readability filter and the de-prioritization of constructions. Furthermore, this is the first search engine for German functional illiteracy contexts. Due to our incorporation of important insights from SLA research, KANSAS is also suited for the use in GSL contexts.

Our next steps include to further refine KANSAS's performance and to conduct more elaborate evaluation studies for both algorithms. Furthermore, we are currently conducting usability studies in which teaching practitioners from the fields of adult literacy and GSL acquisition are evaluating KANSAS in terms of its suitability for real-life use.

Acknowledgments

We are grateful to our project partners Theresa Geppert, Hannes Schröter, and Josef Schrader of the German Institute for Adult Education – Leibniz Centre for Lifelong Learning (DIE) for their valuable collaboration. We also thank the anonymous reviewers for their insightful suggestions.

KANSAS is a research and development project funded by the Federal Ministry of Education and Research (BMBF) as part of the *AlphaDekade*[11] [grant number W143500].

[10] https://cran.r-project.org/web/packages/irr/

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

References

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

Jonathan Brown and Maxine Eskenazi. 2004. Retrieval of authentic documents for reader-specific lexical practice. In *InSTIL/ICALL Symposium 2004*.

Tim Vor der Brück, Sven Hartrumpf, and Hermann Helbig. 2008. A readability checker with supervised learning using deep syntactic and semantic indicators. *Informatica*, 32(4):429–435.

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

Xiaobin Chen and Detmar Meurers. 2017. 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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

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.

Scott A. Crossley, Jerry Greenfield, and Danielle S. Mcnamara. 2008. Assessing text readability using cognitively based indices. *TESOL Quarterly*, 42(3):475–493.

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.

William H. DuBay. 2006. *The Classic Readability Studies*. Impact Information, Costa Mesa, California.

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.

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. 2009. Automatic readability assessment for people with intellectual disabilities. In *ACM SIGACCESS accessibility and computing*, volume 93, pages 84–91.

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.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 276–284.

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*.

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):205–212.

Silke Gausche, Anne Haase, and Diana Zimper. 2014. *Lesen. DVV-Rahmencurriculum*, 1 edition. Deutscher Volkshochschul-Verband e.V., Bonn.

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.

Kevin A. Hallgren. 2012. Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23–34.

Julia Hancke, Detmar Meurers, and Sowmya Vajjala. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on*

Computational Linguistics (COLING), pages 1063–1080, Mumbay, India.

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.

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.

Stephen Krashen. 1977. Some issues relating to the monitor model. *On Tesol*, 77(144-158).

Rudolf Kretschmann and Petra 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 *Proceedings of the fifth international conference on Language Resources and Evaluation*, pages 2231–2234, Genoa, Italy. European Language Resources Association (ELRA).

Luís Marujo, José Lopes, Nuno Mamede, Isabel Trancoso, Juan Pino, Maxine Eskenazi, Jorge Baptista, and Céu Viana. 2009. Porting reap to european portuguese. In *International Workshop on Speech and Language Technology in Education*.

Jacob E. McCarthy and Sarah J. Swierenga. 2010. What we know about dyslexia and web accessibility: a research review. *Universal Access in the Information Society*, 9(2):147–152.

Detmar Meurers, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf, and Niels Ott. 2010. Enhancing authentic web pages for language learners. In *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 10–18, Los Angeles. ACL.

Eleni Miltsakaki and Audrey Troutt. 2007. Read-x: Automatic evaluation of reading difficulty of web text. In *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, pages 7280–7286. Association for the Advancement of Computing in Education (AACE).

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.

Niels Ott and Detmar Meurers. 2011. Information retrieval for education: Making search engines language aware. *Themes in Science and Technology Education*, 3(1-2):9–30.

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.

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.

Wibke Riekmann and Anke Grotlüschen. 2011. Konservative Entscheidungen: Größenordnung des funktionalen Analphabetismus in Deutschland. *REPORT - Zeitschrift für Weiterbildungsforschung*, 3:24–35.

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 232–241. Springer-Verlag New York, Inc.

Richard W. Schmidt. 1990. The role of consciousness in second language learning. *Applied Linguistics*, 11:206–226.

Michael Sharwood Smith. 1993. Input enhancement in instructed SLA. *Studies in Second Language Acquisition*, 15(2):165–179.

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.

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.

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

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 2018)*, Santa Fe, New Mexico, USA. International Committee on Computational Linguistic.

Angelika Wöllstein. 2014. *Topologisches Satzmodell*, 2 edition. Winter, Heidelberg.

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

Victoria Yaneva, Irina Temnikova, and Ruslan Mitkov. 2015. Accessible texts for autism: An eye-tracking study. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, pages 49–57. ACM.

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

A Appendices

↓A\H→	$\alpha 3$	$\alpha 4$	$\alpha 5$	$\alpha 6$	above α
$\alpha 3$	22	7	0	0	0
$\alpha 4$	3	4	0	2	0
$\alpha 5$	0	0	10	0	0
$\alpha 6$	0	0	1	6	3
above α	0	0	2	0	8

Table 3: Raw annotation counts for readability assessment performance pilot (A: algorithm; H: human).

```java
/**
* Assign Alpha readability level given computed features
*
* @return DocumentReadabilityLevel The document's Alpha readability level
*/
public DocumentReadabilityLevel computeReadabilityLevel() {
        if (wordsPerSentence <= 10
                && nSentences <= 5
                && syllablesPerToken <= 3
                && pastPerfectsPerFiniteVerb == 0
                && future1sPerFiniteVerb == 0
                && future2sPerFiniteVerb == 0
                && depClausesPerSentence <= 0.5
                && presentPerfectsPerFiniteVerb <= 0.5
                && typesFoundInSubtlexPerLexicalType >= 0.95) {

                        alphaLevel = LEVEL_3;

        } else if (wordsPerSentence <= 10
                && nSentences <= 10
                && syllablesPerToken <= 5
                && pastPerfectsPerFiniteVerb == 0
                && future1sPerFiniteVerb == 0
                && future2sPerFiniteVerb == 0) {

                        alphaLevel = LEVEL_4;

        } else if (wordsPerSentence <= 12
                && nSentences <= 15
                && pastPerfectsPerFiniteVerb == 0) {

                        alphaLevel = LEVEL_5;

        } else if (wordsPerSentence <= 12
                && nSentences <= 20) {

                        alphaLevel = LEVEL_6;

        } else {

                        alphaLevel = LEVEL_N;

        }

        return alphaLevel;
}
```

Figure 3: A Java code snippet of the algorithm that assigns Alpha readability levels to texts given features such as the number of words per sentence or the number of syllables per token.