THEMES IN SCIENCE AND TECHNOLOGY EDUCATION
Volume 3, Number 1-2, Pages 9-30

Information retrieval for education: making search engines language aware

Niels Ott, Detmar Meurers
{nott,dm}@sfs.uni-tuebingen.de
Universität Tübingen, Germany

Abstract

Search engines have been a major factor in making the web the successful and widely used information source it is today. Generally speaking, they make it possible to retrieve web pages on a topic specified by the keywords entered by the user. Yet web searching currently does not take into account which of the search results are comprehensible for a given user – an issue of particular relevance when considering students in an educational setting. And current search engines do not support teachers in searching for language properties relevant for selecting texts appropriate for language students at different stages in the second language acquisition process.

At the same time, raising language awareness is a major focus in second language acquisition research and foreign language teaching practice, and research since the 20s has tried to identify indicators predicting which texts are comprehensible for readers at a particular level of ability. For example, the military has been interested in ensuring that workers at a given level of education can understand the manuals they need to read in order to perform their job. We present a new search engine approach which makes it possible for teachers to search for texts both in terms of contents and in terms of their reading difficulty and other language properties. The implemented prototype builds on state-of-the art information retrieval technology and exemplifies how a range of readability measures can be integrated in a modular fashion.

Introduction

The Web is a huge repository of information, estimated at over one trillion publicly accessible web pages[1] of which at least 20 billion are estimated to be indexed my the

[1] http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

major search engines[2]. While some pages primarily contain pictures, music, or videos, most of the information on web pages is encoded in natural language, with English being the most commonly used one[3]. In order to provide access to this wealth of information, commercial search engines such as Google, Yahoo, Bing or Ask index the web and support searches for keywords and phrases which result in ranked lists of links to web pages on the corresponding topics.

Yet apart from which words or phrases appear on a web page and what language it is written in, the search engines do not support users in searching for language properties as such, i.e., they are not transparently language aware[4]. At the same time, in an education context, in particular in foreign language teaching, it is crucial to be able to access texts which are written at the right level for the students, which showcase the language patterns which are supposed to be taught next and avoid those lexical and structural patterns which have not yet been taught. This issue is especially important given the need to focus on those language aspects which are teachable at a given point in the language acquisition process (Pienemann 1989). In a related vein, second language acquisition research since the 90s has emphasized the importance of supporting awareness of language categories and forms (cf. Lightbown & Spada 1999) which makes it particularly relevant to be able to obtain texts which include the language properties for which the learners' awareness is to be raised.

Since the web search engines do not index web pages by such language properties, a teacher cannot search for web pages which contain both the content of interest to the students and at the same time prioritize the language properties which are best suited for the students at their particular stage of development in the language acquisition process. When trying to obtain web pages which are appropriate for a particular class level, the only option is to go to particular web sites dedicated to offering material by grade level, such as BBC Bitesize[5] or web sites of educational classroom magazines such as Weekly Reader[6].

[2] http://www.worldwidewebsize.com/

[3] http://www.trnmag.com/Stories/2001/112101/English_could_snowball_on_Net_112101.html

[4] Search engines often include some special treatment of named entities and an analysis of words into stems. This is independent of the main point here, which is that search engines do not support searching for language properties other than the words or phrases entered as search terms and the language a document is written in.

[5] http://www.bbc.co.uk/schools/bitesize

[6] http://www.weeklyreader.com

The situation is reminiscent of the early days of the Web before search engines had taken hold and information was frequently accessed by browsing hierarchically organized directories of information in a top-down, concept-driven manner. While Web directories still exist today, efforts such as the Open Directory Project admit that it is impossible for professional editorial staff to keep up with the growth of the web, and the quality and comprehensiveness of directories suffers accordingly (unless a large community effort of net-citizens steps in as voluntary editors to preform the required manual classification)[7]. For up-to-date access to information on the web, most users of the web today therefore rely on one of the commercial search engines providing the direct, bottom-up access from the words entered as search terms to the pages on the corresponding topics.

In addition to the limited language awareness of the index of the current commercial search engines, which means that a direct search for documents meeting the search criteria relevant for language (and other) educators is not supported, it is useful to revisit the nature of the measures used by the search engines to *rank the search results* presented to the user. Ranking algorithms, such as the well-known PageRank (cf. Langville & Meyer, 2006), which is one of the factors behind Google's success, compute the relative importance of a web page based on the relations the page has with other pages. For the educational need we are focusing on in this paper, such criteria seem less relevant. When trying to obtain texts meeting the needs of language learners, the appropriateness of the language forms and categories for the particular subpopulation of web users is more important for ranking results than the overall 'popularity' of the page.

In this paper, we show how the analysis of language properties can be combined with current approaches to information retrieval in order to obtain a search engine which supports searches for particular contents at particular levels of reading difficulty and showcasing particular language features. We provide an overview of potentially relevant language properties and explore readability measures in more detail. While we focus on readability for the concrete example discussed in this paper, the architecture of our approach is modular enough to support a wide range of language properties needed by language teachers to select appropriate texts for their classes.

The paper is structured as follows: Firstly, we take a closer look at language properties which are relevant in an educational context for indexing web documents and following zoom in on our main example, readability measures. In the next section, we turn to the question how the results of such language property measures

[7] http://dmoz.org/about.html

can be integrated into an information retrieval setup and introduce general text models for that purpose. Finally, we then discuss an experiment showcasing how readability measures can be used to classify and retrieve classes of web pages and evaluate the accuracy of several such methods.

Relevant Language Properties and How to Identify Them

In the introduction, we established the main goal of this paper to develop a search engine capable of indexing language properties which are directly relevant in an educational context. This includes teachers wanting to obtain appropriate documents for their classes as well as learners searching for texts for themselves, be it in a school setting or as part of the voluntary, self-motivated pursuit of knowledge nowadays often referred to as lifelong learning.

General language properties: One language aspect that is of general relevance in an educational context (and beyond) is the *readability* of a text. Readability here relates to a wide range of language properties making a text easy or hard to grasp. Extra-linguistic factors affecting reading, such as the legibility of handwriting or the layout of a text, are usually excluded from readability research (Klare, 1963, p. 1). Readability depends not only on the text but also on the reader. *Text difficulty* can be seen as a more precise term focusing on properties of texts under investigation. *Reading proficiency* is the corresponding concept from the reader's perspective. The two are interrelated: the more proficient readers, the less readable a text needs to be in order to be comprehensible to them. In the next section, we take a closer look at readability and how it can be measured.

Language properties sequenced in language teaching and learning: In the context of language teaching there is a large, second set of language properties that is directly relevant to finding texts which are well-suited for a given class or a particular student. In foreign language education essentially all properties of the language to be learned are sequenced in some way. Typically the orders found in textbooks arise from the pedagogical approach and foreign language teaching tradition. Ideally they are also informed by the study of general cognitive development and the increasing knowledge about developmental stages in second language acquisition (e.g., Pienemann, 1998) – even if the relation between foreign language teaching and second language acquisition research is a complex one (Ellis, 1997a; b).

Vocabulary: At the most basic level, this includes the *vocabulary*, where a teacher will typically want to select texts which practice the vocabulary already introduced while at the same time limiting the number of words which are yet unknown (and, e.g., in lower level classes need to be provided with a definition to be understandable for the student). The issue is related to the effort of dictionary publishers to write the definition in their directories using a defined basic vocabulary, e.g., the 3000 words used in the definitions of the Longman Dictionary of Contemporary English (DCE).

Such grouping of vocabulary into broad groups is taken into account in the so-called Lexical Frequency Profiles (Laufer & Nation, 1995), which are discussed in connection with readability in the next section. However, the order in which particular sets of vocabulary items are taught is partly idiosyncratic and, e.g., dependent on the particular textbook series used. Accordingly, to support ranking the results of a web search in terms of the vocabulary known by a particular class or student, every page must be indexed by the search engine with the ratio of lexical items on the page compared to the list of known vocabulary items at particular cutoffs (i.e., the vocabulary list accumulated from the start up to each chapter for a given textbook series).

Beyond the lexical level: In foreign language teaching, there are typical sequences in which grammar topics and other language patterns beyond the lexical level are introduced. To be able to retrieve appropriate texts, first one has to determine the language properties which are part of a given stage in such a sequence, and second one needs to be able to automatically detect the occurrences of these properties so that they can be indexed by the search engine. The most straightforward way of determining the language properties to be used is to directly refer to those introduced in a given textbook. On that basis, one can then design an NLP approach capable of automatically identifying these textbook-sequenced language patterns. For example, Ott & Ziai (2008) describe an approach based on Constraint Grammar (Karlsson, 1990; Bick, 2001) which identifies *-ing* forms of English verbs and classifies them into the classes taught in typical EFL textbooks: *going-to* future, participles, progressive forms and gerunds.

On the other hand, it would be attractive to replace the reference to textbook-sequenced language patterns with language properties that are directly related to the actual second language acquisition process. If one can identify language properties which are teachable at a given point in the language acquisition process (Pienemann 1989), indexing these properties may also be useful for identifying texts that are appropriate for learners at a particular stage. When assuming that learner perception develops in sync with their production abilities, one can also transfer the methods for assessing the complexity of learner productions to the analysis of texts which are appropriate for learners of a given level. For example, Lu (2009) shows how the *Revised D-Level-Scale* (Covington et al., 2006) encoding the order in which children acquire sentences of increasing complexity can be measured automatically. Related measures could also be taken from the automated analysis of the complexity in second language writing (Lu, 2010). In the same vein, Hawkins & Buttery (2009) identify so-called *criterial features* distinguishing different proficiency levels, which could potentially be repurposed to distinguish texts of different levels and index them accordingly.

Language properties for language awareness: Complementing the need to access texts which are appropriate for particular stages in the language teaching and learning process, there is an important need for texts highlighting particular language properties in the context of raising language awareness.

Research in second language acquisition since the 90s has shown that awareness of language categories and forms is an important ingredient for successful second language acquisition (Long, 1996; Long & Robinson, 1998; Lightbown & Spada, 1999). A wide range of linguistic features can be relevant for language awareness, including morphological, syntactic, semantic, and pragmatic information (Schmidt 1995, p. 30). Especially important in our context is that awareness without input is generally not considered to be sufficient, i.e., it is not enough to learn about linguistic rules and language patterns, but the learner needs to be exposed to those features in actual text to acquire them. In other words, it is crucial for teachers (or the learners themselves) to be able to obtain texts in support of raising language awareness.

Naturally, in order to rise the language awareness of learners it is not sufficient to merely obtain texts containing or showcasing the classes or patterns that are relevant for a particular group of learners. Obtaining the texts is the first step on which traditional teaching methods using those texts can be based. Beyond traditional teaching, obtaining such texts can be particularly important for *input enhancement* (Sharwood Smith, 1993), a technique used to increase the salience of language patterns to make learners notice them. For example, the web-based tool WERTi (Meurers et al., 2010) visually enhances web pages and automatically generates activities for language patterns which are known to be difficult for learners of English, such as determiners and prepositions, phrasal verbs, the distinction between gerunds and to-infinitives, and *wh*question formation. These language forms and patterns are prime examples for the type of language properties which need to be automatically detected, counted and indexed by a language aware search engine. Such search engines then can support a search for web pages which are of interest from a contents point of view as well as containing the language properties targeted by input enhancement and activity generation.

Measuring Readability

As a concrete example of a property of texts that can serve an important function when searching for documents in an educational context, yet so far is not supported by current search engines, we take a closer look at readability.

Readability measures aim at expressing text difficulty in numbers. Traditionally, many such measures were designed to compute the years in U.S. education a reader must have mastered in order to understand the given text. Since the development of the early classical readability formulas in the 1920s, countless new measures have emerged (DuBay, 2004).

The classical formulas mostly make use of surface features of a text, such as the average word length and the average sentence length. Figure 1 shows the well-known Flesch Reading Ease (FRE, Flesch 1948) as an illustration of a traditional readability measure. The FRE has been designed to compute text difficulty on a scale from 0 (very hard) to 100 (very easy).

$$\text{ReadingEase} = 206.356 - 84.6 \cdot \text{AWL}_s - 1.015 \cdot \text{ASL}$$

where

$$\text{AWL}_s = \frac{\text{Number of Syllables}}{\text{Number of Words}}$$ Average word length counted in syllables

$$\text{ASL} = \frac{\text{Number of Words}}{\text{Number of Sentences}}$$ Average sentence length

Figure 1. Example readability measure: Flesch Reading Ease (Flesch 1948)

Lexical factors: At the *lexical level*, the use of the average word length (AWL_S) in the formula in Figure 1 encodes the heuristic that the longer a word is, the more difficult it is likely to be. On the one hand, longer words can encode more complex forms and meanings. On the other, since the early findings of Zipf (1936) it has repeatedly been suggested that longer words are less frequent in language – and infrequent words are more likely to be unknown to readers or language learners.

Other readability measures, such as the one by Dale & Chall (1948a), make use of a word list indicating 'easy' or common words. While this seems to be the more direct way of addressing word frequency, such lists are clearly dependent on genre and topic. Word length as an abstraction of frequency in the above mentioned sense is not affected by this issue. Lexical Frequency Profiles (LFPs) were designed by Laufer &

Nation (1995) for the purpose of measuring the active vocabulary of learners. Laufer & Nation claim that their measure is superior to many others, such as the popular lexical variation measure also known as type/token ratio. An LFP is the outcome of comparing a sample of a learner's writing with three word lists. For example, the *Range* tool by Paul Nation[8] uses the first 1,000 and the second 1,000 words from the General Service List (West 1953) as basis for identifying the most frequent words. Furthermore it uses the New Academic Word List (Coxhead, 2000). A discussion of a variety of word lists for predicting the levels of essays written by second language learners can be found in Pendar & Chapelle (2008).

Clearly, the active and the passive vocabulary of learners are related. Therefore, it should be possible to use LFPs also as a measure for texts which are supposed to be read by learners. However, Bennöhr (2007) discusses the development of a readability formula for assessing the difficulty level in relation to an English as a Foreign Language (EFL) curriculum and remarks that in her experiments she found the contribution of word frequency lists to be an insignificant variable. It is likely that this is due to the fact that the vocabulary used in the EFL curricula is less based on general word frequency and more dependent on the choice of topics in text books.

Syntactic complexity: Returning to the readability formula of Figure 1, the second empirical variable is the average sentence length (ASL), which essentially is used to encode that longer sentences are supposed to be more difficult. Sentence length here can be seen as an approximation of *syntactic complexity*: the longer a sentence is, the more likely it is to contain embedded phrases making it harder to understand.

Bennöhr (2007) presents a readability formula for assessing the difficulty level in relation to an EFL curriculum that uses not only sentence length and word length but also the number of easy and difficult conjunctions as specified in two lists. While still being surface-based in nature, her approximation of sentence complexity is one step closer to analyzing sentence structure by distinguishing different types of embedding indicated by different conjunctions. In recent years more complex statistical classification methods have been proposed (Schwarm & Ostendorf, 2005; Collins-Thompson & Callan, 2005) and work in the Coh-Metrix Project[9] has emphasized the importance of analyzing text cohesion and coherence and of taking a reader's cognitive aptitudes into account for making predictions about reading comprehension (McNamara et al., 2002).

[8] http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

[9] http://cohmetrix.memphis.edu

Relatedly, psycholinguistic research has produced explicit computational models which successfully predict human sentence processing difficulty for a range of sentence types. For example, Boston et al. (2008) present parsing models capable of predicting human reading difficulty. In future work, it will be interesting to explore how such models of human processing difficulty at the sentence level can be integrated into measures of reading difficulty at the text level (e.g., using normalized average or maximum complexity). Instead of merely encoding heuristic surface correlations, such an approach would have the advantage of being rooted in a model of the actual cognitive processes underlying human sentences processing, and could thereby evolve in sync with the evolving understanding of human cognition.

Language Properties meet Information Retrieval

Text and query models for modular integration of language properties

In order to make language properties in text accessible to a search engine, we need an efficient way to store and query them, i.e., for each text we need to store a 'language profile' as part of the search engine model of the text. We tackle this issue by constructing *text models*. These are simple key-value tables containing names of summary measures and their corresponding numeric values. In this paper, we use readability measures to illustrate the approach, but the text models can in principle hold any kind of language properties or other information about documents as long as a numeric value can be determined for that property for every document.

For example, it is possible to encode the text length, the type/token ratio, the number or ratio of words from a certain word list found in the text, or the ratio of gerunds to all verb forms in the text. For each document, our prototype system stores a text model containing the outcome of all analyses modules, in addition to the normal token index of ordinary search engines. An example text model containing some general information about a document as well as readability scores is shown in Table 1.

Document classification takes place at query time using *query models*. A query model is a set of constraints that operate on entries in the text model. It defines possible range constraints for readability measures and other values. To match a query model, all range constraints must be satisfied by the text model of a document; alternatively, some of the constraints can be used to rank the results. In addition, documents must, of course, also match the regular query terms specified by the user.

Table 1. Example text model containing readability scores and generic document information

Key	Value
Generic_AllCharCount	2,904.00
Generic_SentenceCount	161.00
Generic_TokenCount	519.00
R_ARI	3.46
R_ColemanLiau	2.37
R_FleschKincaid	2.65
R_FleschReadingEase	80.77
R_FogIndex	7.86
R_LIX	31.71
R_SMOG	6.95
R_oldDaleChall	7.67

To support easy reference to particular constraints or combination thereof, one can define templates, which can then be referred to in the user interface. Table 2 shows an example query model that could be used to obtain texts of a given length and level, where the templates ‘Long Text’ and ‘Key Stage 4 and above’ are encoded as range queries in terms of the sentence count and the Simple Measure of Gobbledygook (SMOG, McLaughlin, 1969) readability measure.

Classifying documents in terms of such predefined template classes can be done during indexing. The benefit of doing so would be that it does not affect processing time of queries to the search engine. However, it requires re-indexing the entire document collection whenever the classification scheme changes. Using query models, we can modify the functionality of the search engine by simply modifying a part of the user interface code. Re-indexing the potentially huge subset of the web crawled for our search engine then is only needed if the output of new analysis modules is required for a query.

Table 2. Example query model for ‘Long Text’ and ‘Key Stage 4 and above’

Key	Range
Generic_SentenceCount	[150, 1500]
R_SMOG	[7.67, 12]

The LAWSE prototype

We implemented a fully-functional prototype called Language-Aware Search Engine (LAWSE) to realize and test the approach proposed in this paper. State-of-the-art information retrieval technology provided by the Lucene search engine API (Gospodnetić & Hatcher, 2005) was used as the basis for indexing and searching documents for their contents, allowing us to focus on the integration of the language properties.

Our components used for natural language processing (NLP) are hosted in a pipeline based on the Unstructured Information Management Architecture (UIMA, Ferrucci & Lally, 2004). The pre-processing pipeline is built upon standard NLP components. The Java Text Categorization Library[10] is used to ensure that only English documents are being processed; other languages can naturally be added, provided that language resources and NLP components for those languages are available for any analyses requiring language specific processing. SentParBreaker[11] comes to use for splitting the input text into sentences. Tokenizing and part-of-speech tagging components are taken from the OpenNLP[12] project, using the pre-trained statistical models. Syllable counting is provided by a Java port of a rule-based implementation in Perl by Laura Kassner.

We implemented eight traditional readability measures, namely the original Dale-Chall Score (Dale & Chall, 1948a,b), the Flesch Reading Ease (FRE, Flesch, 1948), the Flesch-Kincaid measure Kincaid et al. (1975), the Gunning Fog Index (Gunning, 1968), the Simple Measure of Gobbledygook (SMOG, McLaughlin, 1969), the Läsbarhetsindex (LIX, Björnsson, 1968), the Automated Readability Index (ARI, Smith & Senter, 1967), and the Coleman-Liau Index (Coleman & Liau, 1975).

While most of the readability formulas were designed for manual analysis, users of such formulas today expect the analyses to be conducted by computer programs. Yet, the underlying assumptions are not always made explicit in the original publications of the formulas. Where they are, they often differ from the assumptions underlying current NLP methods. For example, there are different styles of tokenization. In NLP, tokenizers often split contractions such as won't into wo#n't. Flesch (1948), on the other hand, advises: "Count contractions and hyphenated words as one word." Such details must be identified for each classical readability formula and the

[10] http://textcat.sourceforge.net

[11] http://text0.mib.man.ac.uk:8080/scottpiao/sent_detector

[12] http://opennlp.sourceforge.net/

implementations must be adapted accordingly.

Often such adaptation is straightforward, but some rules are quite complex to implement. For example, Flesch's approach to counting syllables in expressions such as *1918* makes it necessary to disambiguate between dates (*nineteen-eighteen*) and numbers (*one thousand-nine-hundredeighteen*). Another example is from Dale and Chall (1948b), who for their word list-based measure include a four-pages list of rules on how to perform look-up on the "Dale list of 3000 familiar words". These instructions also require morphological analysis. For example, the word *treat* is listed as an easy word, hence *treating* is supposed to be an easy word as well. This, however, is not supposed to be the case for *treatment*. We tried to follow these rules as closely as possibles by implementing a morphology-aware word list lookup component, for which details are provided in Ott (2009, p. 53ff).

While we tried to implement the readability measures as close as possible to the originally published approaches, there are two aspects where our implementation diverges from the originals. First, we did not implement the context-dependent counting of syllables in dates versus in numbers. Second, we disregarded the specification of Gunning (1968) not to count inflectional suffixes as syllables.

As a final issue relating to the automated application of the traditional readability formulas, it is relevant to consider that these formulas have been created by regression to some external reference (such as the readability computed for texts of known difficulty, or the mean school grade level of those who successfully complete a test item based on the given text). As such, the regression essentially compensates for some typical errors of the underlying human analysis. Computer programs, on the other hand, make different and more systematic analysis errors than humans. How much this difference affects the outcome of the formulas is an open question. Using formulas designed for automated analysis such as the Automated Readability Index (Smith & Senter, 1967) avoids this issue.

An experiment

To be able to evaluate the performance of LAWSE, we need a set of independently categorized web sites for learners at different stages. Fortunately, the BBC offers extensive web resources for schools, including the BBC Bitesize website[13] which offers texts on a range of topics (English, Maths and Science) for students at different so-called key stages (KS). The KS1 Bitesize targets 5-7 year olds, KS2 Bitesize 7-11

[13] http://www.bbc.co.uk/schools/bitesize

year olds, KS3 Bitesize 11–14 year olds, and finally, for the fourth key stage (14-16 year olds), the GCSE Bitesize[14]. In our experiment, we downloaded the entire BBC Bitesize site and analyzed it using the indexer of the search engine prototype and its readability modules[15]. Since the KS1 materials amounted to only 25 pages containing no text but Flash animations, we removed them. From the remaining materials, we took a random sample of 350 texts for each of the three remaining key stages (2-4). The resulting 1050 documents were randomly assigned to a test set (20%) and a development set (80%). Each document is annotated with the readability measures as computed by the automatic analysis modules we built into the indexer.

In order to test the automatic classification of the documents into key stages, we investigate the results of the readability formulas mentioned in the previous section. To turn the numeric readability scores into the three key stages (2-4), we need to determine two cut-off points for each readability score. We used the following method to estimate these two cutoff points on the basis of the development set. For each measure, we determine the precision-recall break-even point A for KS2 against KS3 and KS4. Similarly, we determine the precision-recall break-even point B for KS4 against KS2 and KS3. In the 10% interval around point A and and the 10% interval around point B, we examine each combination of points A' and B' by computing its balanced F-measure for classifying this middle range as KS3. We store the points A' and B' with the highest F-measure as the two cutoff points for distinguishing between the three classes.

We used the ROCR package (Sing et al., 2007) of R (R Development Core Team 2009) to visualize the classification performance. Figures 2 and 3 show F-measures for intervals classifying the development data into KS2 and KS4 for each readability measure under investigation.

[14] The General Certificate of Secondary Education (GCSE) is the academic qualification in a specified subject which typically is taken at the end of the fourth key stage

[15] The download was conducted on February 9, 2010 using the wget tool.

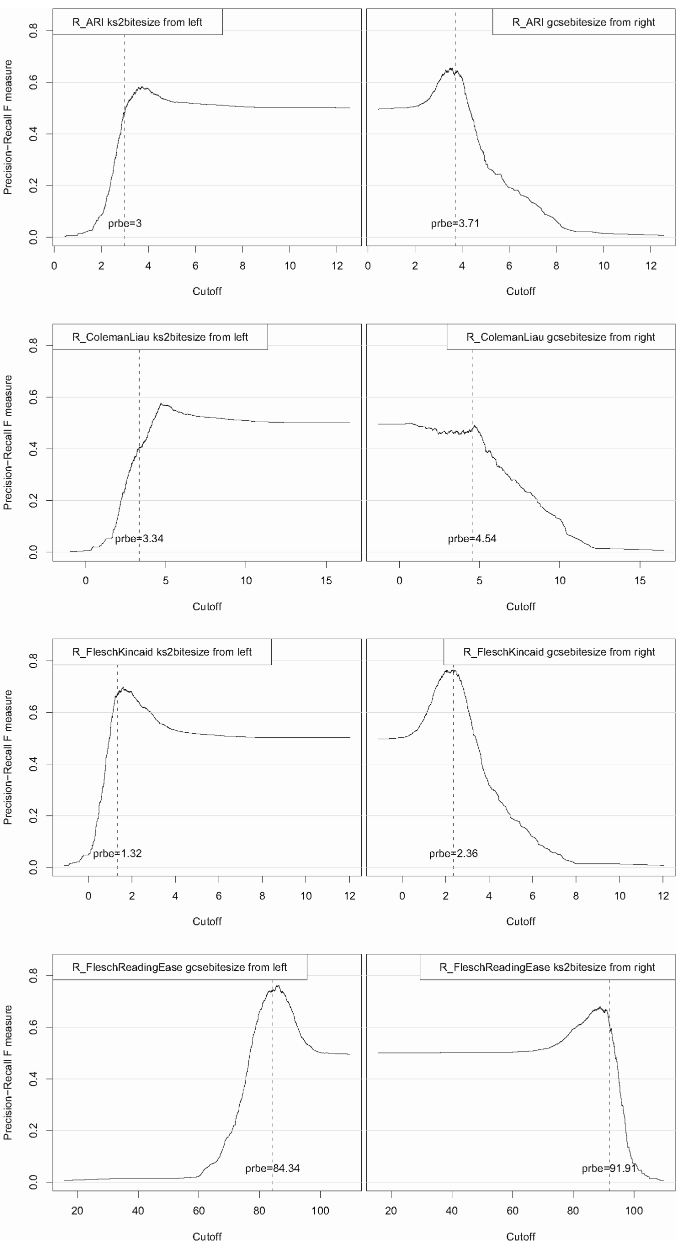

Figure 2. F-measures for intervals ranging from left to a given cutoff (left column) and from a given cutoff to right (right column). Vertical lines mark the precision-recall break-even point.

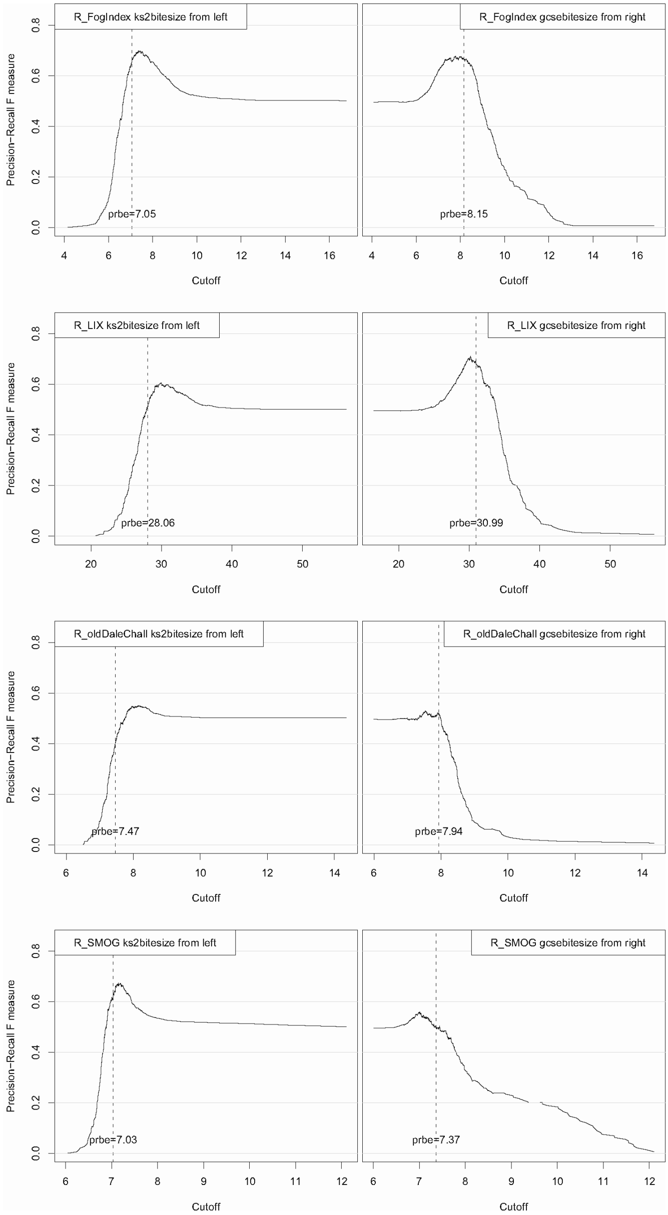

Figure 3. F-measures for intervals ranging from left to a given cutoff (left column) and from a given cutoff to right (right column). Vertical lines mark the precision-recall break-even point.

The intervals for KS2 range from the left side of the plot, moving further up the readability scale. Similarly, for KS4 the intervals range from the right side, moving further down the readability scale. For all readability measures under investigation, the F-measure starts increasing together with the size of the interval since more correctly classified documents are found in the larger interval. However, beyond a certain point, the F-measure decreases since more and more documents of the middle KS3 are wrongly classified as KS2 (or as KS4 in the second binary classification). Note that the Flesch Reading Ease is reversed in comparison to the other readability formulas since it assigns easy documents a high numeric score, not a low one. The precision-recall break-even points serving as the points A and B mentioned above are marked by vertical lines on the plots. We classify the test set based on the cutoff values obtained for the development set. Documents with values indicating a text difficulty below A' are classified as KS2, those with values above B' are classified as KS4, and texts between A' and B' as KS3. Based on this, we computed the accuracy for each readability measure. The estimated cutoff points as well as the classification accuracy computed on the test set are shown in Table 3.

Overall, the classification quality of these traditional readability measures for the BBC Bitesize corpus is quite low. The word list based Dale Chall-Measure measure appears to be the least suitable. The measures that score best in our experiment are the Simple Measure of Gobbledygook (SMOG), the Flesch Reading Ease (FRE) and the Flesch-Kincaid measure. The latter is based on the same analyses as the FRE, mapping the reading ease to U.S. grade levels.

Table 3. Classification accuracy on the test set

Measure	Accuracy	Cutoff *A'*	Cutoff *B'*
R_oldDaleChall	0.34	7.13	8.28
R_ColemanLiau	0.36	2.46	5.44
R_ARI	0.40	2.43	4.08
R_FogIndex	0.42	6.42	8.65
R_LIX	0.46	26.27	32.9
R_SMOG	0.54	7.12	7.67
R_FleschReadingEase	0.54	82.78	94.22
R_FleschKincaid	0.55	1.08	2.47

The goal of this section was to show how the search engine architecture proposed in this paper can provide direct access to texts based on their readability, as an example of a wide range of language properties which are relevant for educators searching for adequate reading materials. The classification of documents using value ranges of readability measures is what the search engine prototype implementation can make use of at query time. Hence, the conducted experiment corresponds to testing the classification ability based on readability measures of the prototype ignoring key word queries. The low accuracy results of the traditional readability measures suggest the use of other readability measures to be integrated as part of the text models – and indeed more complex statistical methods have been proposed for this purpose (Schwarm & Ostendorf, 2005; Collins-Thompson & Callan, 2005), with some recent measures being specifically geared towards evaluating readability in a language learning context (Ozasa et al., 2008).

Related work

The use of readability classification in a search context is reminiscent of the REAP project[16] (Heilman et al. 2008). REAP is a tool providing learners with documents for lexical practice depending on the learners level and needs, and the project has developed sophisticated readability measures (Heilman et al., 2007). Different from our approach, the system collects a corpus of documents off-line and then provides access to the text (i.e., not the original web-page). The pedagogical focus is on lexical practice as the language feature targeted.

A second, related project is Read-X (Miltsakaki & Troutt, 2008), which focuses on providing access to web pages at different levels of reading difficulty. In a first step, the system makes use of a commercial search engine to obtain documents of interest given a particular query entered by a user. All the documents obtained then are downloaded on the fly and filtered for readability. While the use of readability as a post-search filter might be a viable strategy for this text property, such a generate and test strategy is different from and incompatible with our goal of making web pages searchable based on a wide range of language properties relevant in a language teaching context.

Bennöhr (2007) presents Textfinder, a framework that uses a dedicated readability formula for estimating the reading difficulty of documents. Learners initially are requested to submit a piece of their own writing to the system. The learner's reading

[16] http://reap.cs.cmu.edu

level is estimated using the readability formula applied to the learner's writing sample adjusted by a factor. The texts retrieved are presented via a reading interface that asks the learner to rate the difficulty of each text in order to update the learner profile that is used in further queries. It would be interesting to explore how such a feedback loop could be used with the general text and query models of our approach providing access to language properties that can go beyond the scope of readability formulas.

An open issue that remains for all these approaches including ours is whether there are enough texts at each of the reading levels (and with the targeted language properties) on topics of interest to the learner. The fact that there are a number of sites dedicated to offering simplified English texts (Weekly Reader, Simple English Wikipedia, BBC Bitesize, etc.) could be taken to indicate that the general web only contains few such texts – or it could be that appropriate texts are hard to find given that there are no search engines supporting searches for complexity, making it is worthwhile to offer dedicated sites for such texts. If a lack of web pages for lower reading levels turns out to be an issue, an interesting avenue for research would be to actively produce simpler texts in place of just searching for them. Such an approach could also have the advantage of being able to provide the same contents at multiple levels of difficulty. Automated simplification would be most useful to make information accessible, e.g., for people with a medical disability (Carroll et al., 1998), limited education, or when people want to access information in a second language outside of an educational context (e.g., as immigrants). On the other hand, computer tools can support semi-automatic simplification, e.g., by helping teachers simplify and transform a text into a form that is appropriate for a particular purpose, such as fitting it into a sequence of materials for a given class.

Conclusion

We have argued in this paper that current web search engines do not adequately address the needs which arise in a teaching context, especially in foreign language teaching. Yet, the Web offers a rich tapestry of web pages which can be particularly useful as a source of up-to-date and multimedia-enhanced texts. We therefore proposed an extension of web search engines to support retrieval of web pages which on the one hand discuss a topic of interest but at the same time satisfy constraints on the nature of the language.

Depending on the educational context, such constraints can make reference to general properties, such as the readability of a text, or more specific language properties such as the vocabulary used in relation to that covered in a textbook, or the occurrence of particular language patterns targeted by visual input enhancement or automatic activity generation. We have implemented a fully functional prototype using general text and query models and tested it with a range of traditional readability measures. We are currently adding a web crawler to be able to index a wider range of web sites and index them using a variety of language properties, with a particular focus on language categories and patterns of relevance to visual input enhancement and activity generation.

References

Bennöhr, J. (2007). A web-based personalised textfinder for language learners. In G. Rehm, A. Witt, & L. Lemnitzer (Eds.), *Data structures for linguistic resources and applications.* Tübingen: Gunter Narr Verlag.

Bick, E. (2001). The VISL system: Research and applicative aspects of IT-based learning. In *Proceedings of NoDaLiDa, 13th Nordic Conference on Computational Linguistics.* Uppsala, Sweden: Department of Linguistics, Uppsala University. Available from http://beta.visl.sdu.dk/pdf/NoDaLiDa2001.ps.pdf

Björnsson, C.-H. (1968). *Lesbarkeit durch Lix*. Stockholm: Pedagogiskt Centrum i Stockholm.

Boston, M. F., Hale, J. T., Patil, U., Kliegl, R., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, 2(1), 1–12.

Carroll, J., Minnen, G., Canning, Y., Devlin, S., & Tait, J. (1998). Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology.* Madison, Wisconsin: Association for the Advancement of Artificial Intelligence (AAAI).

Coleman, M., & Liau, T. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60, 283–284.

Collins-Thompson, K., & Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13), 1448–1462.

Covington, M. A., He, C., Brown, C., Naçi, L., & Brown, J. (2006). *How complex is that sentence? A proposed revision of the Rosenberg and Abbeduto D-Level Scale* (Computer Analysis of Speech for Psychological Research (CASPR) Research Report No. 2006-01). Athens, GA: The University of Georgia, Artificial Intelligence Center. Retrieved March 10, 2010, from http://www.ai.uga.edu/caspr/2006-01-Covington.pdf

Coxhead, A. (2000). A new academic word list. *Teachers of English to speakers of other languages*, 34(2), 213–238.

Dale, E., & Chall, J. S. (1948a). A formula for predicting readability. *Educational research bulletin; organ of the College of Education*, 27(1), 11–28.

Dale, E., & Chall, J. S. (1948b). A formula for predicting readability: Instructions. *Educational research bulletin; organ of the College of Education*, 27(2), 37–54.

DuBay, W. H. (2004). *The principles of readability*. Costa Mesa, California: Impact Information. Retrieved March 10, 2010, from http://www.impact-information.com/impactinfo/readability02.pdf

Ellis, R. (1997a). SLA and language pedagogy. *Studies in Second Language Acquisition*, 19(1), 69-92.

Ellis, R. (1997b). *SLA research and language teaching*. Oxford: Oxford University Press.

Ferrucci, D., & Lally, A. (2004). UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3–4), 327–348.

Flesch, R. F. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221–233.

Gospodnetić, O., & Hatcher, E. (2005). *Lucene in action*. Greenwich, CT: Manning.

Gunning, R. (1968). *The technique of clear writing* (2nd ed.). New York: McGraw-Hill Book Company.

Hawkins, J. A., & Buttery, P. (2009). Using learner language from corpora to profile levels of proficiency – Insights from the English Profile Programme. In *Studies in language testing: The social and educational impact of language assessment.* Cambridge: Cambridge University Press.

Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL'07)* (pp. 460–467). Rochester, New York: Association for Computational Linguistics. Retrieved March 10, 2010, from http://aclweb.org/anthology-new/N07-1058

Heilman, M., Zhao, L., Pino, J., & Eskenazi, M. (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* (pp. 80–88). Columbus, Ohio: Association for Computational Linguistics. Retrieved March 10, 2010, from http://aclweb.org/anthology-new/W08-0910

Karlsson, F. (1990). Constraint grammar as a framework for parsing running text. In *Proceedings of the 13th Conference on Computational Linguistics (COLING), Volume 3* (pp. 168–173). Helsinki, Finland: Association for Computational Linguistics. Retrieved March 10, 2010, from http://aclweb.org/anthology-new/C90-3030

Kincaid, J. P., Fishburne, R. P. J., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease formula) for Navy enlisted personnel* (Research Branch Report No. 8-75). Millington, TN: Naval Technical Training Command.

Klare, G. R. (1963). *The measurement of readability*. Ames, Iowa: Iowa State University Press.

Langville, A. N., & Meyer, C. D. (2006). *Google's pagerank and beyond: The science of search engine rankings*. Princeton, N.J: Princeton University Press.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3), 307–322.

Lightbown, P. M., & Spada, N. (1999). *How languages are learned*. Oxford: Oxford University Press.

Long, M. H. (1996). The role of linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). New York: Academic Press.

Long, M. H., & Robinson, P. (1998). Focus on form: Theory, research, and practice. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 15–41). Cambridge: Cambridge University Press.

Lu, X. (2009). Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, 14(1), 3-28.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15.

McLaughlin, G. H. (1969). SMOG grading – a new readability formula. *Journal of Reading*, *12*(8), 639–646. Retrieved March 10, 2010, from http://webpages.charter.net/ghal/SMOG_Readability_Formula-G._Harry_McLaughlin_(1969).pdf

McNamara, D. S., Louwerse, M. M., & Graesser, A. C. (2002). *Coh-metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension.* Proposal of Project funded by the Office of Educational Research and Improvement, Reading Program. Available from http://cohmetrix.memphis.edu/cohmetrixpr/archive/Coh-MetrixGrant.pdf

Meurers, D., Ziai, R., Amaral, L., Boyd, A., Dimitrov, A., Metcalf, V., Ott, N. (2010). Enhancing authentic web pages for language learners. In *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-5) at NAACL-HLT 2010.* Los Angeles: Association for Computational Linguistics. Retrieved March 10, 2010, from http://purl.org/dm/papers/meurers-ziai-et-al-10.html

Miltsakaki, E., & Troutt, A. (2008). Real time web text classification and analysis of reading difficulty. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08* (pp. 89–97). Columbus, Ohio: Association for Computational Linguistics. Available from http://aclweb.org/anthology-new/W08-0911

Ott, N. (2009). *Information retrieval for language learning: An exploration of text difficulty measures*. ISCL master's thesis, Universität Tübingen, Seminar für Sprachwissenschaft, Tübingen, Germany. Available from http://drni.de/zap/ma-thesis

Ott, N., & Ziai, R. (2008). *ICALL activities for gerunds vs. to-infinitives: A constraint-grammarbased extension to the New WERTi system.* Term paper for seminar *Using Natural Language Processing to Foster Language Awareness in Second Language Learning*, Universität Tübingen. Available from http://drni.de/zap/werti-gerunds

Ozasa, T., Weir, G., & Fukui, M. (2008). Toward a readability index for Japanese learners of EFL. In *Proceedings of the 13th Conference of Pan-Pacific Association of Applied Linguistics (PAAL'08).* University of Hawaii, Manoa: Pan-Pacific Association of Applied Linguistics. Retrieved March 10, 2010, from http://www.cis.strath.ac.uk/cis/research/publications /papers/strath_cis_publication_2263.pdf

Pendar, N., & Chapelle, C. (2008). Investigating the promise of learner corpora: Methodological issues. *CALICO Journal*, 25(2), 189-206. Available from https://calico.org/html/article_689.pdf

Pienemann, M. (1989). Is Language Teachable? Psycholinguistic Experiments and Hypotheses. *Applied Linguistics*, 10(1), 52-79.

Pienemann, M. (1998). *Language processing and second language development: Processability theory*. Amsterdam: John Benjamins.

R-Development Core Team (2009). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from http://www.R-project.org

Schmidt, R. (1995). Consciousness and foreign language: A tutorial on the role of attention and awareness in learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 1–63). Honolulu: University of Hawaii Press.

Schwarm, S., & Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)* (pp. 523–530). Ann Arbor, Michigan: Association for Computational Linguistics. Available from http://aclweb.org/anthology-new/P05-1065

Sharwood Smith, M. (1993). Input enhancement in instructed SLA: Theoretical bases. *Studies in Second Language Acquisition*, 15, 165-179.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2007). ROCR: Visualizing the performance of scoring classifiers [Computer software manual]. Available from http://rocr .bioinf.mpi-sb.mpg.de (R package version 1.0-2)

Smith, E. A., & Senter, R. J. (1967). *Automated Readability Index* (Tech. Rep. No. AMRL-TR-66-220). Wright-Patterson Airforce Base, OH: Aerospace Medical Research Laboratories.

West, M. (1953). *A general service list of English words*. London: Longmans. Zipf, G. K. (1936). *The psycho-biology of language*. London: Routledge.