Exploring Measures of "Readability" for Spoken Language: Analyzing linguistic features of subtitles to identify age-specific TV programs

Sowmya Vajjala and **Detmar Meurers**
LEAD Graduate School, Department of Linguistics
University of Tübingen
{sowmya,dm}@sfs.uni-tuebingen.de

Abstract

We investigate whether measures of readability can be used to identify age-specific TV programs. Based on a corpus of BBC TV subtitles, we employ a range of linguistic readability features motivated by Second Language Acquisition and Psycholinguistics research.

Our hypothesis that such readability features can successfully distinguish between spoken language targeting different age groups is fully confirmed. The classifiers we trained on the basis of these readability features achieve a classification accuracy of 95.9%. Investigating several feature subsets, we show that the authentic material targeting specific age groups exhibits a broad range of linguistics and psycholinguistic characteristics that are indicative of the complexity of the language used.

1 Introduction

Reading, listening, and watching television programs are all ways to obtain information partly encoded in language. Just like books are written for different target groups, current TV programs target particular audiences, which differ in their interests and ability to understand language. For books and text in general, a wide range of readability measures have been developed to determine for which audience the information encoded in the language used is accessible. Different audiences are commonly distinguished in terms of the age or school level targeted by a given text.

While for TV programs the nature of the interaction between the audio-visual presentation and the language used is a relevant factor, in this paper we want to explore whether the language by itself is equally characteristic of the particular age groups targeted by a given TV program. We thus focused on the language content of the program as encoded in TV subtitles and explored the role of text complexity in predicting the intended age group of the different programs.

The paper is organized as follows. Section 2 introduces the corpus we used, and section 3 the readability features employed and their motivation. Section 4 discusses the experimental setup, the experiments we conducted and their results. Section 5 puts our research into the context of related work, before section 6 concludes and provides pointers to future research directions.

2 Corpus

The BBC started subtitling all the scheduled programs on its main channels in 2008, implementing UK regulations designed to help the hearing impaired. Van Heuven et al. (2014) constructed a corpus of subtitles from the programs run by nine TV channels of the BBC, collected over a period of three years, January 2010 to December 2012. They used this corpus to compile an English word frequencies database SUBTLEX-UK[1], as a part of the British Lexicon Project (Keuleers et al., 2012). The subtitles of four channels (CBeebies, CBBC, BBC News and BBC Parliament) were annotated with the channel names.

While CBeebies targets children aged under 6 years, CBBC telecasts programs for children 6–12 years old. The other two channels (News, Parliament) are not assigned to a specific age-group, but it seems safe to assume that they target a broader, adult audience. In sum, we used the BBC subtitle corpus with a three-way categorization: CBeebies, CBBC, Adults.

Table 1 shows the basic statistics for the overall corpus. For our machine learning experiments, we use a balanced subcorpus with 3776 instances for each class. As shown in the table, the programs for

[1] http://crr.ugent.be/archives/1423

Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) @ EACL 2014, pages 21–29,
Gothenburg, Sweden, April 26-30 2014.

Program Category	Age group	# texts	avg. tokens per text	avg. sentence length (in words)
CBEEBIES	$<$ 6 years	4846	1144	4.9
CBBC	6–12 years	4840	2710	6.7
Adults (News + Parliament)	$>$ 12 years	3776	4182	12.9

Table 1: BBC Subtitles Corpus Description

the older age-groups tend to be longer (i.e., more words per text) and have longer sentences. While text length and sentence length seem to constitute informative features for predicting the age-group, we hypothesized that other linguistic properties of the language used may be at least as informative as those superficial (and easily manipulated) properties. Hence, we explored a broad linguistic feature set encoding various aspects of complexity.

3 Features

The feature set we experimented with consists of 152 lexical and syntactic features that are primarily derived from the research on text complexity in Second Language Acquisition (SLA) and Psycholinguistics. There are four types of features:

Lexical richness features (LEX): This group consists of various part-of-speech (POS) tag densities, lexical richness features from SLA research, and the average number of senses per word.

Concretely, the POS tag features are: the proportion of words belonging to different parts of speech (nouns, proper nouns, pronouns, determiners, adjectives, verbs, adverbs, conjunctions, interjections, and prepositions) and different verb forms (VBG, VBD, VBN, VBP in the Penn Treebank tagset; Santorini 1990) per document.

The SLA-based lexical richness features we used are: type-token ratio and corrected type-token ratio, lexical density, ratio of nouns, verbs, adjectives and adverbs to the number of lexical words in a document, as described in Lu (2012).

The POS information required to extract these features was obtained using Stanford Tagger (Toutanova et al., 2003). The average number of senses for a non-function word was obtained by using the MIT WordNet API[2] (Finlayson, 2014).

Syntactic complexity features (SYNTAX): This group of features encodes the syntactic complexity of a text derived from the constituent structure of the sentences. Some of these features are derived from SLA research (Lu, 2010), specifically: mean lengths of production units (sentence, clause, t-unit), sentence complexity ratio (# clauses/sentence), subordination in a sentence (# clauses per t-unit, # complex t-units per t-unit, # dependent clauses per clause and t-unit), co-ordination in a sentence (# co-ordinate phrases per clause and t-unit, # t-units/sentence), and specific syntactic structures (# complex nominals per clause and t-unit, # VP per t-unit). Other syntactic complexity features we made use of are the number of NPs, VPs, PPs, and SBARs per sentence and their average length (in terms of # words), the average parse tree height and the average number of constituents per sub-tree.

All of these features were extracted using the Berkeley Parser (Petrov and Klein, 2007) and the Tregex pattern matcher (Levy and Andrew, 2006).

While the selection of features for these two classes is based on Vajjala and Meurers (2012), for the following two sets of features, we explored further information available through psycholinguistic resources.

Psycholinguistic features (PSYCH): This group of features includes an encoding of the average Age-of-acquisition (AoA) of words according to different norms as provided by Kuperman et al. (2012), including their own AoA rating obtained through crowd sourcing. It also includes measures of word familiarity, concreteness, imageability, meaningfulness and AoA as assigned in the MRC Psycholinguistic database[3] (Wilson, 1988). For each feature, the value per text we computed is the average of the values for all the words in the text that had an entry in the database.

While these measures were not developed with readability analysis in mind, we came across one paper using such features as measures of word difficulty in an approach to lexical simplification (Jauhar and Specia, 2012).

[2] http://projects.csail.mit.edu/jwi

[3] http://www.psych.rl.ac.uk/

Celex features (CELEX): The Celex lexical database (Baayen et al., 1995) for English consists of annotations for the morphological, syntactic, orthographic and phonological properties for more than 50k words and lemmas. We included all the morphological and syntactic properties that were encoded using character or numeric codes in our feature set. We did not use frequency information from this database.

In all, this feature set consists of 35 morphological and 49 syntactic properties per lemma. The set includes: proportion of morphologically complex words, attributive nouns, predicative adjectives, etc. in the text. A detailed description of all the properties of the words and lemmas in this database can be found in the Celex English Linguistic Guide[4].

For both the PSYCH and CELEX features, we encode the average value for a given text. Words which were not included in the respective databases were ignored for this computation. On average, around 40% of the words from texts for covered by CELEX, 75% by Kuperman et al. (2012) and 77% by the MRC database.

We do not use any features encoding the occurrence or frequency of specific words or n-grams in a document.

4 Experiments and Results

4.1 Experimental Setup

We used the WEKA toolkit (Hall et al., 2009) to perform our classification experiments and evaluated the classification accuracy using 10-fold cross validation. As classification algorithm, we used the Sequential Minimal Optimization (SMO) implementation in WEKA, which marginally outperformed (1–1.5%) some other classification algorithms (J48 Decision tree, Logistic Regression and Random Forest) we tried in initial experiments.

4.2 Classification accuracy with various feature groups

We discussed in the context of Table 1 that sentence length may be a good surface indicator of the age-group. So, we first constructed a classification model with only one feature. This yielded a classification accuracy of 71.4%, which we consider as our baseline (instead of a basic random baseline of 33%).

[4] http://catalog.ldc.upenn.edu/docs/LDC96L14/eug_a4.pdf

We then constructed a model with all the features we introduced in section 3. This model achieves a classification accuracy of 95.9%, which is a 23.7% improvement over the sentence length baseline in terms of classification accuracy.

In order to understand what features contribute the most to classification accuracy, we applied feature selection on the entire set, using two algorithms available in WEKA, which differ in the way they select feature subsets:

- *InfoGainAttributeEval* evaluates the features individually based on their Information Gain (IG) with respect to the class.
- *CfsSubsetEval* (Hall, 1999) chooses a feature subset considering the correlations between features in addition to their predictive power.

Both feature selection algorithms use methods that are independent of the classification algorithm as such to select the feature subsets.

Information Gain-based feature selection results in a ranked list of features, which are independent of each other. The Top-10 features according to this algorithm are listed in Table 2.

Feature	Group
avg. AoA (Kuperman et al., 2012)	PSYCH
avg. # PPs in a sentence	SYNTAX
avg. # instances where the lemma has stem and affix	CELEX
– avg. parse tree height	SYNTAX
– avg. # NPs in a sentence	SYNTAX
avg. # instances of affix substitution	CELEX
– avg. # prep. in a sentence	LEX
avg. # instances where a lemma is not a count noun	CELEX
avg. # clauses per sentence	SYNTAX
– sentence length	SYNTAX

Table 2: Ranked list of Top-10 features using IG

As is clear from their description, all Top-10 features encode different linguistic aspects of a text. While there are more syntactic features followed by Celex features in these Top-10 features, the most predictive feature is a psycholinguistic feature encoding the average age of acquisition of words. A classifier using only the Top-10 IG features achieves an accuracy of 84.5%.

Applying CfsSubsetEval to these Top-10 features set selects the six features not prefixed by a

hyphen in the table, indicating that these features do not correlate with each other (much). A classifier using only this subset of 6 features achieves an accuracy of 84.1%.

We also explored the use of CfsSubsetEval feature selection on the entire feature set instead of using only the Top 10 features. From the total of 152 features, CfsSubsetEval selected a set of 41 features. Building a classification model with only these features resulted in a classification accuracy of 93.9% which is only 2% less than the model including all the features.

Table 3 shows the specific feature subset selected by the CfsSubsetEval method, including some examples illustrating the morphological features. The method does not provide a ranked list, so the features here simply appear in the order in which they are included in the feature vector.

preposition phrases
t-units
co-ordinate phrases per t-unit
lexical words in total words
interjections
conjunctive phrases
word senses
verbs
verbs, past participle (VBN)
proper nouns
plural nouns
avg. corrected type-token ratio
avg. AoA acc. to ratings of Kuperman et al. (2012)
avg. AoA acc. to ratings of Cortese and Khanna (2008)
avg. word imageability rating (MRC)
avg. AoA according to MRC
morph. complex words (e.g., *sandbank*)
morph. conversion (e.g., *abandon*)
morph. irrelevant (e.g., *meow*)
morph. obscure (e.g., *dedicate*)
morph. may include root (e.g., *imprimatur*)
foreign words (e.g., *eureka*)
words with multiple analyses (e.g., *treasurer*)
noun verb affix compounds (e.g., *stockholder*)
lemmas with stem and affix (e.g., *abundant=abound+ant*)
flectional forms (e.g., *bagpipes*)
clipping allomorphy (e.g., *phone* vs. *telephone*)
deriv. allomorphy (e.g., *clarify–clarification*)
flectional allomorphy (e.g., verb *bear* $\mapsto$ adjective *born*)
conversion allomorphy (e.g., *halve–half*)
lemmas with affix substitution (e.g., *active=action+ive*)
words with reversion (e.g., *downpour*)
uncountable nouns
collective, countable nouns
collective, uncountable nouns
post positive nouns.
verb, expression (e.g., *bell the cat*)
adverb, expression (e.g., *run amok*)
reflexive pronouns
wh pronouns
determinative pronouns

Table 3: CfsSubsetEval feature subset

All of these features except for the psycholinguistic features encode the number of occurrences averaged across the text (e.g., average number of prepositions/sentence in a text) unless explicitly stated otherwise. The psycholinguistic features encode the average ratings of words for a given property (e.g., average AoA of words in a text).

Table 4 summarizes the classification accuracies with the different feature subsets seen so far, with the feature count shown in parentheses.

Feature Subset (#)	**Accuracy**	**SD**
All Features (152)	95.9%	0.37
Cfs on all features (41)	93.9%	0.59
Top-10 IG features (10)	84.5%	0.70
Cfs on IG (6)	84.1%	0.55

Table 4: Accuracy with various feature subsets

We performed statistical significance tests between the feature subsets using the Paired T-tester (corrected), provided with WEKA and all the differences in accuracy were found to be statistically significant at $p < 0.001$. We also provide the Standard Deviation (SD) of the test set accuracy in the 10 folds of CV per dataset, to make it possible to compare these experiments with future research on this dataset in terms of statistical significance.

Table 5 presents the classification accuracies of individual features from the Top-10 features list (introduced in Table 2).

Feature	**Accuracy**
AoA_Kup_Lem	82.4%
# pp	74.0%
# stem & affix	77.7%
avg. parse tree height	73.4%
# np	73.0%
# substitution	74.3%
# prep	72.0%
# uncountable nouns	68.3%
# clauses	72.5%
sentence length	71.4%

Table 5: Accuracies of Top-10 individual features

The table shows that all but one of the features individually achieves a classification accuracy above 70%. The first feature (AoA_Kup_Lem)

alone resulted in an accuracy of 82.4%, which is quite close to the accuracy obtained by all the Top-10 features together (84.5%).

To obtain a fuller picture of the impact of different feature groups, we also performed ablation tests removing some groups of features at a time. Table 6 shows the results of these tests along with the SD of the 10 fold CV. All the results that are statistically different at $p < 0.001$ from the model with all features (95.9% accuracy, 0.37 SD) are indicated with a *.

Features	**Acc.**	**SD**
All − AoA_Kup_Lem	95.9%	0.37
All − All AoA Features	95.6%	0.58
All − PSYCH	95.8%	0.31
All − CELEX	94.7%*	0.51
All − CELEX − PSYCH	93.6%*	0.66
All − CELEX − PSYCH − LEX (= SYNTAX only)	77.5%*	0.99
LEX	93.1%*	0.70
CELEX	90.0%*	0.79
PSYCH	84.5%*	1.12

Table 6: Ablation test accuracies

Interestingly, removing the most predictive individual feature (AoA_Kup_Lem) from the feature set did not change the overall classification accuracy at all. Removing all of the AoA features or all of the psycholinguistic features also resulted in only a very small drop. The combination of the linguistic features, covering lexical and syntactic characteristics as well as the morphological, syntactic, orthographic, and phonological properties from Celex, thus seem to be equally characteristic of the texts targeting different age-groups as the psycholinguistic properties, even though the features are quite different in nature.

In terms of separate groups of features, syntactic features alone performed the worst (77.5%) and lexical richness features the best (93.1%).

To investigate which classes were mixed up by the classifier, consider Table 7 showing the confusion matrix for the model with all features on a 10-fold CV experiment.

We find that CBeebies is more often confused with the CBBC program for older children (156+214) and very rarely with the program for adults (1+2). The older children programs (CBBC) are more commonly confused with programs for adults (36+58) compared to CBeebies (1+2), which is expected given that the CBBC audience is closer in age to adults than the CBeebies audience.

classified as →	**CBeebies**	**CBBC**	**Adults**
CBeebies (0–6)	3619	156	1
CBBC (6–12)	214	3526	36
Adults (12+)	2	58	3716

Table 7: Confusion Matrix

Summing up, we can conclude from these experiments that the classification of transcripts into age groups can be informed by a wide range of linguistics and psycholinguistic features. While for some practical tasks a few features may be enough to obtain a classification of sufficient accuracy, the more general take-home message is that authentic texts targeting specific age groups exhibit a broad range of linguistics characteristics that are indicative of the complexity of the language used.

4.3 Effect of text size and training data size

When we first introduced the properties of the corpus in Table 1, it appeared that sentence length and the overall text length could be important predictors of the target age-groups. However, the list of Top-10 features based on information gain was dominated by more linguistically oriented syntactic and psycholinguistic features.

Sentence length was only the tenth best feature by information gain and did not figure at all in the 43 features chosen by the CfsSubsetEval method selecting features that are highly correlated with the class prediction while having low correlation between themselves. As mentioned above, sentence length as an individual feature only achieved a classification accuracy of 71.4%.

The text length is not a part of any feature set we used, but considering the global corpus properties we wanted to verify how well it would perform and thus trained a model with only text length (#sentences per text) as a feature. This achieved a classification accuracy of only 56.7%.

The corpus consists of transcripts of whole TV programs and hence an individual transcript text typically is longer than the texts commonly used in readability classification experiments. This raises the question whether the high classification accuracies we obtained are the consequences of the larger text size.

As a second issue, the training size available for the 10-fold cross-validation experiments is com-

paratively large, given the 3776 text per level available in the overall corpus. We thus also wanted to study the impact of the training size on the classification accuracy achieved.

Pulling these threads together, we compared the classification accuracy against text length and training set size to better understand their impact. For this, we trained models with different text sizes (by considering the first 25%, 50%, 75% or 100% of the sentences from each text) and with different training set sizes (from 10% to 100%).

Figure 1 presents the resulting classification accuracy in relation to training set size for the different text sizes. All models were trained with the full feature set (152 features), using 10-fold cross-validation as before.

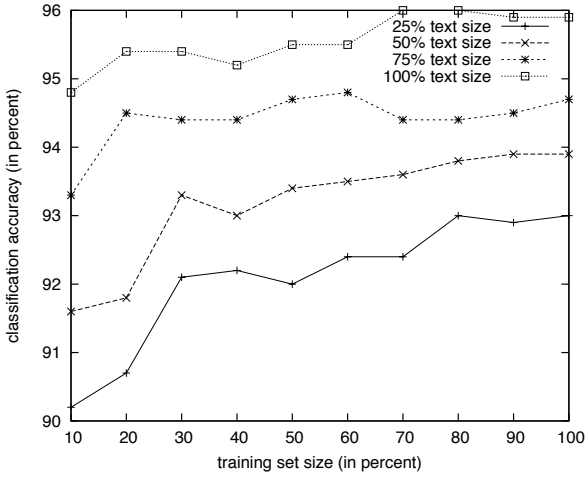

Figure 1: Classification accuracy for different text sizes and training set sizes

As expected, both the training set size and the text size affect the classification accuracy. However, the classification accuracy even for the smallest text and training set size is always above 90%, which means that the unusually large text and training size is not the main factor behind the very high accuracy rates.

In all four cases of text size, there was a small effect of training set size on the classification accuracy. But the effect reduced as the text size increased. At 25% text size, for example, the classification accuracy ranged 90–93% (mean 92.1%, SD 0.9) as the training set size increased from 10% to 100%. However, at 100% text size, the range was only 94.8–96% (mean 95.6%, SD 0.4).

Comparing the results in terms of text size alone, larger text size resulted in better classification accuracy in all cases, irrespective of the training set size. A longer text will simply provide more information for the various linguistic features, enabling the model to deliver better judgments about the text. However, despite the text length being reduced to one fourth of its size, the models built with our feature set always collect enough information to ensure a classification accuracy of at least 90%.

In the above experiments, we varied the text size from 10% to 100%. But since these are percentages, texts from CBBC and Adults on average still are longer than CBEEBIES texts. While this reflects the fact that TV transcripts in real life are of different length, we also wanted to see what happens when we eliminate such length differences.

We thus trained classification models fixing the length of all documents to a concrete absolute length, starting from 100 words (rounded off to the nearest sentence boundary) increasing the text size until we achieve the best overall performance. Figure 2 displays the classification accuracy we obtained for the different (maximum) text sizes, for all features and feature subsets.

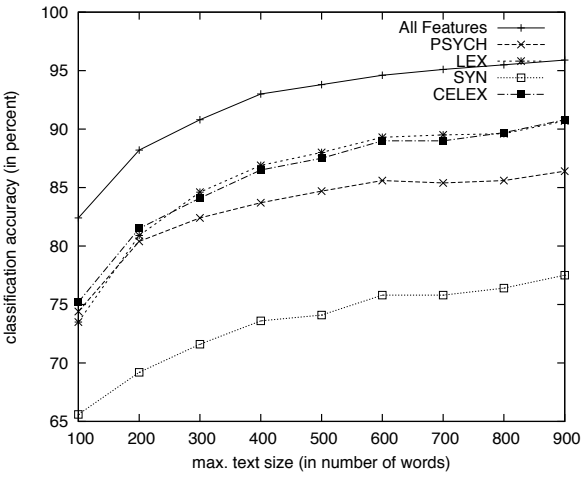

Figure 2: Classification accuracy for different absolute text sizes (in words)

The plot shows that the classification accuracy already reaches 80% accuracy for short texts, 100 words in length, for the model with all features. It rises to above 90% for texts which are 300 words long and reaches the best overall accuracy of almost 96% for texts which are 900 words in length. All the feature subsets too follow the same trend, with varying degrees of accuracy that is always lower than the model with all features.

While in this paper, we focus on documents, the issue whether the data can be reduced further

to perform readability at the sentence level is discussed in Vajjala and Meurers (2014a).

5 Related Work

Analyzing the complexity of written texts and choosing suitable texts for various target groups including children is widely studied in computational linguistics. Some of the popular approaches include the use of language models and machine learning approaches (e.g., Collins-Thompson and Callan, 2005; Feng, 2010). Web-based tools such as REAP[5] and TextEvaluator[6] are some examples of real-life applications for selecting English texts by grade level.

In terms of analyzing spoken language, research in language assessment has analyzed spoken transcripts in terms of syntactic complexity (Chen and Zechner, 2011) and other textual characteristics (Crossley and McNamara, 2013).

In the domain of readability assessment, the Common Core Standards (`http://www.corestandards.org`) guideline texts were used as a standard test set in the recent past (Nelson et al., 2012; Flor et al., 2013). This test set contains some transcribed speech. However, to the best of our knowledge, the process of selecting suitable TV programs for children as explored in this paper has not been considered as a case of readability assessment of spoken language before.

Subtitle corpora have been created and used in computational linguistics for various purposes. Some of them include video classification (Katsiouli et al., 2007), machine translation (Petukhova et al., 2012), and simplification for deaf people (Daelemans et al., 2004). But, we are not aware of any such subtitle research studying the problem of automatically identifying TV programs for various age-groups.

This paper thus can be seen as connecting several threads of research, from the analysis of text complexity and readability, via the research on measuring SLA proficiency that many of the linguistic features we used stem from, to the computational analysis of speech as encoded in subtitles. The range of linguistic characteristics which turn out to be relevant and the very high precision with which the age-group classification can be performed, even when restricting the input to artificially shortened transcripts, confirm the usefulness of connecting these research threads.

[5] `http://reap.cs.cmu.edu`

[6] `https://texteval-pilot.ets.org/TextEvaluator`

6 Conclusions

In this paper, we described a classification approach identifying TV programs for different age-groups based on a range of linguistically-motivated features derived from research on text readability, proficiency in SLA, and psycholinguistic research. Using a collection of subtitle documents classified into three groups based on the targeted age-group, we explored different classification models with our feature set.

The experiments showed that our linguistically motivated features perform very well, achieving a classification accuracy of 95.9% (section 4.2). Apart from the entire feature set, we also experimented with small groups of features by applying feature selection algorithms. As it turns out, the single most predictive feature was the age-of-acquisition feature of Kuperman et al. (2012), with an accuracy of 82.4%. Yet when this feature is removed from the overall feature set, the classification accuracy is not reduced, highlighting that such age-group classification is informed by a range of different characteristics, not just a single, dominating one. Authentic texts targeting specific age groups exhibit a broad range of linguistics and psycholinguistic characteristics that are indicative of the complexity of the language used.

While an information gain-based feature subset consisting of 10 features resulted in an accuracy of 84.5%, a feature set chosen using the CfsSubsetEval method in WEKA gave an accuracy of 93.9%. Any of the feature groups we tested exceeded the random baseline (33%) and a baseline using the popular sentence length feature (71.4%) by a large margin. Individual feature groups also performed well at over 90% accurately in most of the cases. The analysis thus supports multiple, equally valid perspectives on a given text, each view encoding a different linguistic aspect.

Apart from the features explored, we also studied the effect of the training set size and the length of the text considered for feature extraction on classification accuracy (Section 4.3). The size of training set mattered more when the text size was smaller. Text size, which did not work well as an individual feature, clearly influences classification accuracy by providing more information for model building and testing.

In terms of the practical relevance of the results, one question that needs some attention is how well the features and trained models generalize across different type of TV programs or languages. While we have not yet investigated this for TV subtitles, in experiments investigating the cross-corpus performance of a model using the same feature set, we found that the approach performs well for a range of corpora composed of reading materials for language learners (Vajjala and Meurers, 2014b). The very high classification accuracies of the experiments we presented in the current paper thus seem to support the assumption that the approach can be useful in practice for automatically identifying TV programs for viewers of different age groups.

Regarding the three class distinctions and the classifier setup we used in this paper, the approach can also be generalized to other scales and a regression setup (Vajjala and Meurers, 2013).

6.1 Outlook

The current work focused mostly on modeling and studying different feature groups in terms of their classification accuracy. Performing error analysis and looking at the texts where the approach failed may yield further insights into the problem. Some aspects of the text that we did not consider include discourse coherence or topic effects. Studying these two aspects can provide more insights into the nature of the language used in TV programs directed at viewers of different ages. A cross-genre evaluation between written and spoken language complexity across age-groups could also be insightful.

On the technical side, it would also be useful to explore the possibility of using a parser tuned to spoken language, to check if this helps improve the classification accuracy of syntactic features.

While in this paper we focused on English, a related readability model also performed well for German (Hancke et al., 2012) so that we expect the general approach to be applicable to other languages, subject to the availability of the relevant resources and tools.

Acknowledgements

We would like to thank Marc Brysbaert and his colleagues for making their excellent resources available to the research community. We also thank the anonymous reviewers for their useful feedback. This research was funded by LEAD Graduate School (GSC 1028, `http://purl.org/lead`), a project of the Excellence Initiative of the German federal and state governments.

References

Harald R. Baayen, Richard Piepenbrock, and Leon Gulikers. 1995. The CELEX lexical database. `http://catalog.ldc.upenn.edu/LDC96L14`.

Maio Chen and Klaus Zechner. 2011. Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 722–731, Portland, Oregon, June.

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

Michael J. Cortese and Maya M. Khanna. 2008. Age of acquisition ratings for 3,000 monosyllabic words. *Behavior Research Methods*, 43:791–794.

Scott Crossley and Danielle McNamara. 2013. Applications of text analysis tools for spoken response grading. *Language Learning & Technology*, 17:171–192.

Walter Daelemans, Anja Hoethker, and Erik F. Tjong Kim Sang. 2004. Automatic sentence simplification for subtitling in Dutch and English. In *Fourth International Conference on Language Resources And Evaluation (LREC)*, pages 1045–1048.

Lijun Feng. 2010. *Automatic Readability Assessment*. Ph.D. thesis, City University of New York (CUNY).

Mark Alan Finlayson. 2014. Java libraries for accessing the princeton wordnet: Comparison and evaluation. In *Proceedings of the 7th Global Wordnet Conference*, pages 78–85.

Michael Flor, Beata Beigman Klebanov, and Kathleen M. Sheehan. 2013. Lexical tightness and text complexity. In *Proceedings of the Second Workshop on Natural Language Processing for Improving Textual Accessibility (PITR) held at ACL*, pages 29–38, Sofia, Bulgaria. ACL.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The weka data mining software: An update. *The SIGKDD Explorations*, 11:10–18.

Mark A. Hall. 1999. *Correlation-based Feature Selection for Machine Learning*. Ph.D. thesis, The University of Waikato, Hamilton, NewZealand.

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): Technical Papers*, pages 1063–1080, Mumbai, India.

Sujay Kumar Jauhar and Lucia Specia. 2012. Uowshef: Simplex – lexical simplicity ranking based on contextual and psycholinguistic features. In *In proceedings of the First Joint Conference on Lexical and Computational Semantics (SEM)*.

Polyxeni Katsiouli, Vassileios Tsetsos, and Stathes Hadjiefthymiades. 2007. Semantic video classification based on subtitles and domain terminologies. In *Proceedings of the 1st International Workshop on Knowledge Acquisition from Multimedia Content (KAMC)*.

Emmanuel Keuleers, Paula Lacey, Kathleen Rastle, and Marc Brysbaert. 2012. The british lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic english words. *Behavior Research Methods*, 44:287–304.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44(4):978–990.

Roger Levy and Galen Andrew. 2006. Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *5th International Conference on Language Resources and Evaluation*, pages 2231–2234, Genoa, Italy. European Language Resources Association (ELRA).

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Languages Journal*, pages 190–208.

Jessica Nelson, Charles Perfetti, David Liben, and Meredith Liben. 2012. Measures of text difficulty: Testing their predictive value for grade levels and student performance. Technical report, The Council of Chief State School Officers.

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York, April.

Volha Petukhova, Rodrigo Agerri, Mark Fishel, Yota Georgakopoulou, Sergio Penkale, Arantza del Pozo, Mirjam Sepesy Maucec, Martin Volk, and Andy Way. 2012. Sumat: Data collection and parallel corpus compilation for machine translation of subtitles. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 21–28, Istanbul, Turkey. European Language Resources Association (ELRA).

Beatrice Santorini. 1990. Part-of-speech tagging guidelines for the Penn Treebank, 3rd revision, 2nd printing. Technical report, Department of Computer Science, University of Pennsylvania.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-ofspeech tagging with a cyclic dependency network. In *HLT-NAACL*, pages 252–259, Edmonton, Canada.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *In Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA) at NAACL-HLT*, 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 Natural Language Processing for Improving Textual Accessibility (PITR) held at ACL*, pages 59—-68, Sofia, Bulgaria. ACL.

Sowmya Vajjala and Detmar Meurers. 2014a. Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. ACL.

Sowmya Vajjala and Detmar Meurers. 2014b. Readability assessment for text simplification: From analyzing documents to identifying sentential simplifications. *International Journal of Applied Linguistics, Special Issue on Current Research in Readability and Text Simplification, edited by Thomas François and Delphine Bernhard*.

Walter J.B. Van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. Subtlex-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, pages 1–15.

Michael D. Wilson. 1988. The mrc psycholinguistic database: Machine readable dictionary, version 2. *Behavioural Research Methods, Instruments and Computers*, 20(1):6–11.