Towards grounding computational linguistic approaches to readability: Modeling reader-text interaction for easy and difficult texts

Sowmya Vajjala
Iowa State University, USA
sowmya@iastate.edu

Detmar Meurers
LEAD Graduate School and Research Network
University of Tübingen, Germany
dm@sfs.uni-tuebingen.de

Alexander Eitel
University of Freiburg, Germany
alexander.eitel@psychologie.uni-freiburg.de

Katharina Scheiter
LEAD Graduate School and Research Network
Leibniz-Institut für Wissensmedien (IWM), Tübingen, Germany
k.scheiter@iwm-tuebingen.de

Abstract

Computational approaches to readability assessment are generally built and evaluated using gold standard corpora labeled by publishers or teachers rather than being grounded in observations about human performance. Considering that both the reading process and the outcome can be observed, there is an empirical wealth that could be used to ground computational analysis of text readability. This will also support explicit readability models connecting text complexity and the reader's language proficiency to the reading process and outcomes.

This paper takes a step in this direction by reporting on an experiment to study how the relation between text complexity and reader's language proficiency affects the reading process and performance outcomes of readers after reading We modeled the reading process using three eye tracking variables: fixation count, average fixation count, and second pass reading duration. Our models for these variables explained 78.9%, 74% and 67.4% variance, respectively. Performance outcome was modeled through recall and comprehension questions, and these models explained 58.9% and 27.6% of the variance, respectively. While the online models give us a better understanding of the cognitive correlates of reading with text complexity and language proficiency, modeling of the offline measures can be particularly relevant for incorporating user aspects into readability models.

1 Introduction

Automatic Readability Assessment (ARA) has been an active area of research in computational linguistics over the past two decades, resulting in a wide range of supervised machine learning models that used both theory driven and data driven features (Petersen, 2007; Feng, 2010; Vajjala and Meurers, 2014b; Jiang et al., 2015, for example). Though the purpose of ARA is to predict text complexity, the eventual goal is ensure that the predictions reflect the comprehension difficulties in the reader. However, so far, ARA models primarily used training corpora that were based on judgements of teachers and other language experts, and not based on the actual reading performance of students, as was also recently criticized by education researchers (Valencia et al., 2014; Williamson et al., 2014; Cunningham and Mesmer, 2014). While this can be considered a shortcoming, obtaining large amounts of data on the actual reading performance of target population is difficult and time consuming. One way to tackle this is to develop a hybrid ARA model, which separately models text complexity and user's language comprehension ability and link them through another model. In this paper, we describe one approach to integrate reader and text characteristics into a single model for automatic readability assessment.

Eye-tracking was employed as a method to understand various NLP problems such as annotation task difficulty (Tomanek et al., 2010; Joshi et al., 2013; Joshi et al., 2014; Barrett and Søgaard, 2015),

Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity,
pages 38–48, Osaka, Japan, December 11-17 2016.

translation difficulty (Mishra et al., 2013), and studying reader eye movements using standard corpora (Martínez-Gómez et al., 2012; Matthies and Søgaard, 2013). Cognitive psychologists have for a long time studied eye-movement patterns of readers to understand the cognitive processes in reading and comprehension, and what causes reading difficulty (Just and Carpenter, 1980; Rayner, 1998; Clifton Jr et al., 2007). Studying the eye movements of readers during reading considering both text and reader factors will give us a better understanding about the online link (during reading) between text complexity and reader proficiency. Asking readers to answer questions about the text will give us an understanding about the offline link (after reading) between text complexity and reader proficiency. Finally, having a means to combine readability models with a model of readers' language proficiency will provide us a solution to create efficient content recommendation system for readers, considering reader characteristics into account.

On this background, we report on an experiment that studies the relation between text complexity and reader proficiency during and after reading. To our knowledge, this is the first reported study to combine online and and offline measures in one experiment, and develop models for more than one form of questions. In sum, the contributions of this paper are:

1. We explored modeling the cognitive correlates of text complexity and reader proficiency by studying the eye movements of readers using three eye-tracking measures: fixation count, average fixation count, and second pass reading durations.

2. We modeled how readers will respond to two types of questions (recall and comprehension) after reading the texts of varying reading difficulty, based on their language proficiency. We believe that this model paves way for the development of better text recommendation systems for readers based on their proficiency and the readability of the text itself.

The paper is organized as follows: Section 2 surveys existing literature on the topic and puts our research in context. Section 3 explains the experimental procedure, Section 4 explains the data analysis methods and variables studied, Section 5 describes the results and Section 6 summarizes the main conclusions of this paper.

2 Related Work

The effect of text complexity on a reader's comprehension was studied in cognitive psychology literature in the 70s and 80s, for various reader groups such as high school students (Evans, 1972), elderly readers (Walmsley et al., 1981) and primary school students (Green and Olsen, 1988; Smith, 1988). The primary conclusion from this research so far has been that carefully written simplified versions of texts resulted in better comprehension. Britton and Gülgöz (1991) showed that rewriting a text based on Kintsch's reading comprehension model (Kintsch and van Dijk, 1978) resulted in better free recall of the text.

Apart from this above mentioned research on complex texts and their revised versions, studying eye movement patterns was shown to be useful in understanding the cognitive processes involved in reading and comprehension (e.g., Just and Carpenter, 1980; Rayner, 1998; Clifton et al., 2007). Eye tracking, though time and cost consuming, provides a more natural way to study the reading processes and allows us to study the processes like re-reading of the text by readers. Eye movements in reading research are typically studied in terms of fixations, saccades and regressions. Fixations refer to the relatively stationary positions of the eye at specific areas of text and saccades refer to the rapid eye movements between fixations. Regressions refer to the cases where the reader revisits and fixates on parts that were already read. Reader's comprehension difficulties were shown to manifest in longer fixations, shorter saccades and more regressions in previous research (cf. Rayner (1998) for a review).

Text readability and its effects on reading comprehension have not been explored much from the perspective of reader proficiency and reading performance, to our knowledge. Two studies that are closely related to the current research are Rayner et al. (2006) and Crossley et al. (2014). Rayner et al. (2006) explicitly studied how text's difficulty level affects eye movement measures in reading and concluded that the text difficulty rating correlated strongly with average fixation duration, number of

fixations and total time. Readers' performance with comprehension questions did not have a significant correlation with text difficulty in their experiment. More recently, Crossley et al. (2014) used a moving window self-paced reading task to study the effect of text simplification on text comprehension and reading time of second language learners of English. The moving window shows a sentence step by step, without showing the full text, and with no means to do re-reading. Comprehension was assessed by means of yes/no questions and the subjects also participated in an English proficiency test. Their results showed that while text complexity affected the reading time, this effect was no longer significant upon including the subject's English reading proficiency as a covariate. In terms of comprehension, while text complexity was significant, the effect of text complexity on comprehension was less for highly proficient readers compared to low proficiency readers. Our study differs from Rayner et al. (2006) in terms of materials and analysis methods. While they used a collection on unrelated text passages for their study, we use same texts written in two versions for the experiments. Our study differs from Crossley et al. (2014) in terms of the experimental methods. While they did a self-paced reading time study with a moving window approach, we used eye-tracking, which allows us to observe more reading variables. Finally, our study differs from both the studies in terms of additional eye-tracking and reader performance variables studied.

3 Experiment

Participants: 48 non-native English speakers studying in a German university participated in this study. Their English proficiency was evaluated using a standardized online c-test (Taylor, 1953) used at the University for placement testing, and the average score of the participants was 72.6 (range: $[21, 112]$) where a score of 100+ is considered highly proficient. The participants came from different L1 backgrounds. We collected this information but it was not used in the analyses reported here.

Texts: Four texts, each written in two versions (advanced and beginner), taken from onestopenglish.com, were used in this study. Texts from the same source were used in related research (Crossley et al., 2014). Since the participants read the text from an eye-tracker, we restricted the length of texts used to 300-350 words in both versions. They read a practice text and answered questions before starting the actual experiment. Eight recall questions and six comprehension questions per text were created, which had the same answer in both versions of a text. While the recall questions primarily dealt with the factual information in the text and had short answers spanning a few words, comprehension questions were yes/no questions that needed drawing inferences. All the authors worked together to create the questions, and the final list of questions was created after a discussion to reach consensus about the questions and answers to the questions.[1] The responses of participants were manually evaluated by a graduate student, by comparing them with the gold standard answers.

Table 1 shows some statistics about the texts used, along with additional information about the complexity of the texts based on automated approaches.

Text_Version	Num. Sentences	NumWords	Flesch-Kincaid	VM	Surprisal
1_Difficult	12	296	14.75	5.2	207.5
1_Easy	15	298	10.09	3.9	147.2
2_Difficult	11	286	11.00	4.2	193.2
2_Easy	14	234	6.30	3.1	112.3
3_Difficult	11	248	11.10	4.1	165.4
3_Easy	13	230	7.74	3.0	124.6
4_Difficult	12	312	13.70	5.4	181.9
4_Easy	14	306	11.08	4.8	144.4

Table 1: Number of words in the texts used for the experiment

[1]The texts in both versions, c-test and the questions asked can be accessed in the Appendix of Vajjala (2015).

Flesch-Kincaid Grade Level (Kincaid et al., 1975) is a standard readability formula. VM refers to the readability score assigned by the model of Vajjala and Meurers (2014a), which is a regression model based on several lexical and syntactic features, and outputs a score between 1–6, with higher values indicating more difficult texts. Surprisal is a psycholinguistic measure of expected cognitive load during sentence processing, based on information theory. We took the average total surprisal for all sentences from Roark parser (Roark et al., 2009) as a measure of surprisal for each text. Though we modeled different notions of complexity, we only report about the models with the binary complexity from onestopenglish.com in this paper.

Procedure: We employed Latin square design for the experiment, making sure each participant read all four texts, alternating between easy and difficult versions. No participant read the same text in two versions. They answered questions on paper after each text and the eye-tracker was re-calibrated for their next reading. Participants were randomly assigned to one of the four experimental conditions, which differed in the order of texts read. We conducted the experiment using iViewX^{TM} Hi-speed eye-tracker from Senso Motoric Instruments (SMI) and collected the reading data through SMI BeGaze[2] software with Reading package.

4 Analysis Methods

4.1 Modeling

We modeled our experimental data using Generalized Additive Mixed Models (GAMMs, Wood (2006)) and in a cross validation setup. GAMMs are a combination of Generalized Additive Models (GAM) and mixed effects models. Whereas GAM allows us to model complex non-linear interactions between variables by modeling the response variable as a function a smoothed version of predictor variables, GAMM adds an additional layer of modeling convenience to GAM by allowing us to delineate between variables with fixed effects and random effects as in a mixed effects model. In these models, fixed effects refer to the independent variables considered in the experiment design and random effect variables are used to model the variation due to sampling choices. In our experiment, texts and participants can be considered random variables, since we cannot sample all possible texts or humans in a single experiment. Following previous research which used GAMMs for linguistic studies (Wieling et al., 2014; Nixon et al., 2015), we constructed our GAMM models as implemented in the mgcv[3] (Wood, 2011) package in R.

4.2 Experimental Variables

Dependent Variables: We report on three eye-tracking variables and two reader performance measures as our dependent variables:[4]

Three eye-tracking measures – average fixation count (average number of times a reader fixates on a word) and average fixation duration (average duration of such fixations in milliseconds), and average second pass reading time (in milliseconds) – were analyzed to study study how the relation between text complexity and reader proficiency affects online processing of these texts. Previous research in cognitive psychology has shown that a reader's comprehension difficulties are reflected in eye-movements through increased (Rayner, 1998) and longer (Just and Carpenter, 1980; Rayner, 1998) fixations. Both these measures are also known to correlate with text difficulty in the experiment described by Rayner et al. (2006).

Two reader performance outcome measures – number of correct answers for recall and comprehension questions – were used as dependent variables related to offline measures. Each text had eight recall and six comprehension questions, which are the maximum scores the participants can get per text respectively.

[2] http://www.smivision.com/en/gaze-and-eye-tracking-systems/products/begaze-analysis-software.html

[3] https://cran.r-project.org/web/packages/mgcv/

[4] We studied other eye-tracking variables as well. More details can be found in Vajjala (2015, ch. 4).

Independent (fixed effect) Variables: We considered the binary text complexity (categorical: elementary and advanced as easy and difficult respectively) and the reader's English proficiency (numeric) as two primary independent variables. Additionally, hypothesizing that there could be some effect of reading texts one after another, we also considered the order in which the participant read a given text (which depends on the experimental condition) as another independent variable.

Random Effects Variables: The two likely random effect factors that can cause a systematic variation in model construction in this experiment are participants and texts. Thus, we considered both of them as random effect variables.

5 Results

For each dependent variable, multiple GAMM models were constructed with different random effect structures, interaction components, and smoothing functions. Model performance was compared in terms of variance explained (R^2) and statistical significance of the differences were compared using the itsadug[5] (van Rij et al., 2016) package in R. We report the results with only the best performing model for each variable below.

Online measures – Fixation count: The best performing model for fixation count explained 78.9% of the variance and included a three way interaction between text difficulty, reader proficiency and text order, modeled with a tensor product smooth function and with log-transformed fixation counts. While the interaction between proficiency and text complexity was not by itself a significant factor in this model, the three way interaction between proficiency, complexity and text order was significant. The model summary, showing the parametric coefficients and the significant smooth terms can be seen in Table 2.

Parametric Coefficients

Variable	Estimate	Std. Err.	t	p-value
Intercept	2.478	0.0481	51.51	< 0.001
Difficulty-Easy	-0.178	0.023	-7.61	$*< 0.001$

Significant Smooth Terms

Variable	RE?	DF	F	p-value
te(Proficiency, Order): Difficult	No	8.095	4.273	$*< 0.001$
te(Proficiency, Order): Easy	No	4.544	7.549	$*< 0.001$
participant	Yes	41.86	11.020	$*< 0.001$
text	Yes	2.154	3.015	*0.007

Variance Explained (R^2 adj): 78.9%

Table 2: Best Performing Model for Fixation Count (* indicates statistically significant)

The negative co-efficient for difficulty in Table 2 shows that the fixation count decreases as one goes from difficult to easy texts. It also shows that the random variations due to the individual differences among participants and texts are both significant factors. This reiterates the usefulness of considering random effects and going beyond linear models, in understanding the relation between eye-tracking variables, reader proficiency, and text complexity. A visualization of the three way interaction between proficiency, text complexity and text order is presented in Figure 1.

We can observe from the figure that low proficiency readers make higher number of fixations (darker color indicates lower values) when they read difficult texts compared to easy texts. However, the number of fixations also increase depending when they read a text. The fixation counts are clearly lower for the texts they read in the early parts of the experiment. However, this effect (and that of text complexity) is less pronounced in more proficient readers. Thus, we can conclude that fixation count is affected by changes in both reader proficiency and text complexity.

[5] https://cran.r-project.org/web/packages/itsadug/

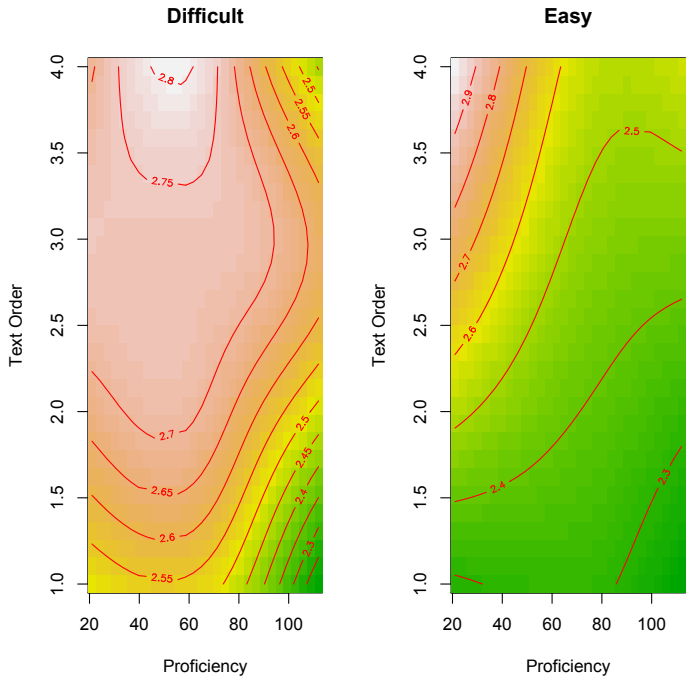

Figure 1: Interaction between Text difficulty, Reader Proficiency and Text Order for fixation count

Online measures – Average Fixation Duration (AFD): The best model for AFD explained 74% of the variance and uses the default thin plate regression spline smoothing without performing any transformations on the AFD. Table 3.

Parametric Coefficients

Variable	Estimate	Std. Err.	t	p-value
Intercept	146.863	10.1432	14.479	< 0.001
Diff-Easy	0.323	4.0968	0.079	0.937
TextOrder	9.0981	2.1122	4.307	*<0.001

Significant Smooth Terms

Variable	RE?	DF	F	p-value
Proficiency	No	2.031	3.121	*0.044
Participant	Yes	39.64	7.43	$<$ *0.001
Text	Yes	2.63	5.80	$<$ *0.001

Variance Explained (R^2 adj): 74%

Table 3: Summary of the GAMM model for Average Fixation Duration

Only proficiency ($p < 0.05$) and text order ($p < 0.001$) had a significant effect for AFD, with higher proficiencies resulting in lower fixation durations. The relationship between proficiency and AFD was non-linear and both the random effects were significant ($p < 0.001$). None of the interactions were significant. These results lead us to a conclusion that AFD is not affected by text complexity, but is affected by a reader's proficiency, in our experimental data.

Online measures – Second pass reading duration: The best performing model for second pass duration explained 67.4% of the variance and included a three way interaction between text difficulty, reader proficiency and text order, modeled with a tensor product smooth function and with log-transformed fixation counts. Table 4 summarizes the coefficients of the GAMM model. As can be observed from the model summary in Table 4, text difficulty, text order, the three way interaction between proficiency, text order and difficulty, and both the random effects – all were significant predictors for this model.

Parametric Coefficients

Variable	Estimate	Std. Error	t value	p-value
Intercept	6.11	0.09	70.43	< 0.001
Difficulty-Easy	-0.296	0.046	-6.395	*< 0.001
TextOrder	0.521	0.026	19.878	* < 0.001

Significant Smooth Terms

Variable	is Random Effect?	Est. deg. of freedom	F	p-value
te(Proficiency,TextOrder):Difficult	No	7.785	22.296	*< 0.001
te(Proficiency,TextOrder):Easy	No	5.32	29.646	*< 0.001
Participant	Yes	38.43	5.456	*< 0.001

Variance Explained (R^2 adj): 67.4%

Table 4: Best Performing Model for Second Pass Duration

Offline measures – Recall: The best performing model involved a three way interaction, as in fixation count and second pass reading duration, and with tensor smooths. Table 5 shows the model summary in terms of its coefficients and smooth terms. As we can see in the parametric coefficients, positive co-efficient for difficulty variable indicates that the performance of participants with recall questions increased as one moved from difficult to easy texts, which means they scored higher for easy texts. There is also a significant interaction between proficiency, text order and text difficulty, and both the random effects were significant. This leads us to a conclusion that the participants' responses to recall questions depends on both text difficulty and reader proficiency, along with other factors.

Parametric Coefficients

Variable	Estimate	Std. Error	t value	p-value
Intercept	3.006	0.321	9.347	< 0.001
Difficulty-Easy	0.679	0.192	3.527	*< 0.001
TextOrder	0.467	0.089	5.202	*< 0.001

Significant Smooth Terms

Variable	is Random Effect?	Est. deg. of freedom	F	p-value
s(Proficiency)	No	0.9887	51.29	*< 0.001
te(Proficiency,TextOrder):Difficulty-Difficult	No	5.78	3.194	*0.006
Participant	Yes	29.272	1.806	*< 0.001
Text	Yes	2.009	3.817	*0.0014

Variance Explained (R^2 adj): 58.9%

Table 5: Best Performing Model for Recall

Offline measures – Comprehension: The best model for comprehension scores explained only 27.6% of variance compared to other variables, with only proficiency being a significant predictor, apart from the random variation due to texts used. Table 6 shows the model summary for comprehension scores. It is interesting to note that text complexity did not affect reader's comprehension of a text. Thus, though we hypothesized that comprehension scores are affected by text complexity, it seems to depend only on the language proficiency of the participant and not on the reading level of the text, as was also shown by Crossley et al. (2014). However, the low performance of this model compared to others described above needs further study, in order to understand what affects readers' performance on such yes/no comprehension questions.

The experiments discussed above demonstrate that the eye-tracking measures we studied seem to be affected by text complexity, proficiency and their interaction. We also observed that one of the outcome variables, recall, seem to be influenced by both text complexity and readers' language proficiency while only the latter affected the comprehension scores.

Parametric Coefficients

Variable	Estimate	Std. Err.	t	p-value
Intercept	3.951	0.279	14.12	< 0.001
Diff:Easy	0.039	0.154	0.255	0.799
TextOrder	0.108	0.077	1.401	0.163

Significant Smooth Terms

Variable	RE?	DF	F	p-value
Proficiency	No	1.313	10.051	< 0.001
Text	Yes	2.39	4.351	0.001

Variance Explained (R^2 adj): 27.6%

Table 6: Summary of the GAMM model for Comprehension Scores

Relation between online and offline measures: Given this background, we briefly explored whether the effect of text complexity and proficiency on online processing can be used to explain the differences in the learning outcomes of the participants. We used mediation analysis as a means to address this question. Mediation analysis is the process of studying the relationship between the dependent and independent variables by means of a third "mediator" variable. In mediation models, it is generally hypothesized that the independent variable influences the mediator, which in turn influences the dependent variable. It is usually used to understand the underlying mechanism behind a known relationship. We performed this analysis using the mediation package in R (Tingley et al., 2014)[6] considering the eye-tracking measures as mediator variables and the recall and comprehension scores as the dependent variables, and text complexity and language proficiency as the independent variables respectively. To perform the mediation analyses, we need to ensure that the relationship between the mediator and the dependent variable is statistically significant in the first place. Among the three eye-tracking measures we explored, only average fixation duration showed a significant correlation with recall and comprehension. So, we performed the mediation analysis only with this as the mediator variable. There was no significant mediation effect of average fixation duration on either recall or comprehension performance of the participants. Thus, we can conclude that eye-tracking is not mediating the participant differences in the recall and comprehension scores.

6 Conclusion

In this paper, we described an approach to model the relation between text complexity and the reader's language proficiency. Our approach has two parts: modeling the cognitive correlates of text complexity using eye tracking, and a modeling for performance outcomes of the reader by asking them to answer questions about the texts they read. These experiments were motivated by the ultimate goal of recommending appropriate texts to readers considering both text complexity and reader proficiency as influencing factors. In terms of the cognitive correlates, while fixation count and second pass duration were affected by both text complexity and reader proficiency, average fixation duration was affected by reader proficiency alone. For performance measures, while the recall model explained 58.9% variance and had both text complexity and reader proficiency as significant predictors, the comprehension model model was affected by proficiency alone, and explained only 27.6% of the variance.

The results from the our analyses support the conclusion that the eye-movement patterns of the readers are sensitive to the complexity of the text they are reading, as was seen by increased fixation counts and second pass reading time with increased text complexity. Average fixation duration was affected by language proficiency but not text complexity. In terms of the outcome measures, on one hand, the performance of recall and comprehension models leaves scope for a lot of improvement to be used in real life application scenarios. But, it also reiterates the importance of considering differences between question types during modeling. Further, our comprehension questions here primarily consisted of Yes/No

[6] http://cran.r-project.org/web/packages/mediation/

questions that relied on short pieces of information. Modeling responses to other questions that require detailed responses, and that address different levels of comprehension (Day and Park, 2005) may help us build better models in future.

The approach described in this experiment used human encoded text complexity labels and an automated proficiency test. Replacing human created labels with an automated readability assessment model prediction will make the offline measures models applicable to new texts, making it useful for text recommendation based on reader language proficiency and text complexity. Thus, the approach can provide a means to personalize text recommendations considering both reading level and reader characteristics into account, without requiring any search logs per user. This approach can also avoid the problem of creating huge amounts of user based reading data to train readability assessment models by keeping the text complexity model separate from the user proficiency model, but combining them together into a ensemble model.

The current paper demonstrates a simple way of combining a model of text complexity and a simple model of reader proficiency to predict the recall and comprehension of a given reader and a given text. However, text complexity is much richer than a single number, as the wide range of linguistic features considered in Vajjala (2015) illustrate, and future modeling of the link between text complexity and reader proficiency arguably should consider incorporating different aspects of language form and content (vocabulary, syntax, discourse coherence, etc.) into the model. Similarly, future modeling of users should integrate more aspects of language proficiency (e.g., complexity, accuracy, fluency), and cognitive individual differences (e.g., working memory capacity) to build a richer proficiency profile for the user. Consequently, a comprehensive combined model of text complexity and reader proficiency will need to consider all these aspects and their potential interaction.

Acknowledgments

We would like to thank the three anonymous reviewers for their comments and Harald Baayen for patiently answering our questions about interpreting GAMM results. This research was funded by the LEAD Graduate School & Research Network [GSC1028], a project of the Excellence Initiative of the German federal and state governments, and received support through grants ANR-11-LABX-0036 (BLRI) and ANR-11-IDEX-0001-02 (A*MIDEX).

References

Maria Barrett and Anders Søgaard. 2015. Reading behavior predicts syntactic categories. *CoNLL 2015*, page 345.

Bruce K. Britton and Sami Gülgöz. 1991. Using kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology*, 83:329–345.

Charles Clifton Jr, Adrian Staub, and Keith Rayner, 2007. *Eye movement research: A window on mind and brain*, chapter Eye movements in reading words and sentences, pages 341–372. Oxford:Elsevier Ltd.

Scott A. Crossley, Hae Sung Yang, and Danielle S. McNamara. 2014. What's so simple about simplified texts? a computational and psycholinguistic investigation of text comprehension and text processing. *Reading in a Foreign Language*, 26(1):92–113.

James W. Cunningham and Heidi Anne Mesmer. 2014. Quantitative measurement of text difficulty: What's the use? *The Elementary School Journal*, 115(2):pp. 255–269.

Richard R. Day and Jeong-Suk Park. 2005. Developing reading comprehension questions. *Reading in a Foreign Language*, 17(1):60–73.

Ronald V. Evans. 1972. The effect of transformational simplification on the reading comprehension of selected high school students. In *Journal of Literacy Research*.

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

Georgia M. Green and Margaret S. Olsen, 1988. *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*, chapter 5. Preferences for and Comprehension of Original and Readability Adapted Materials, pages 115–140. Lawrence Erlbaum Associates.

Zhiwei Jiang, Gang Sun, Qing Gu, Tao Bai, and Daoxu Chen. 2015. A graph-based readability assessment method using word coupling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 411–420, Lisbon, Portugal, September. Association for Computational Linguistics.

Salil Joshi, Diptesh Kanojia, and Pushpak Bhattacharyya. 2013. More than meets the eye: Study of human cognition in sense annotation. In *HLT-NAACL*, pages 733–738.

Aditya Joshi, Abhijit Mishra, Nivvedan Senthamilselvan, and Pushpak Bhattacharyya. 2014. Measuring sentiment annotation complexity of text. In *ACL (2)*, pages 36–41.

M.A. Just and P.A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87:329–355.

J. P. Kincaid, R. P. Jr. Fishburne, R. L. Rogers, and B. 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.

Walter Kintsch and Teun A van Dijk. 1978. Toward a model of text comprehension and productions. *Pschological Review*, 85(5):363–394, September.

Pascual Martínez-Gómez, Tadayoshi Hara, and Akiko N Aizawa. 2012. Recognizing personal characteristics of readers using eye-movements and text features. In *COLING*, pages 1747–1762.

Franz Matthies and Anders Søgaard. 2013. With blinkers on: Robust prediction of eye movements across readers. In *EMNLP*, pages 803–807.

Abhijit Mishra, Pushpak Bhattacharyya, Michael Carl, and IBC CRITT. 2013. Automatically predicting sentence translation difficulty. In *ACL (2)*, pages 346–351.

Jessie S Nixon, Jacolien van Rij, Peggy Mok, Harald Baayen, and Yiya Chen. 2015. Eye movements reflect acoustic cue informativity and statistical noise. *Experimental Linguistics*, page 50.

Sarah E. Petersen. 2007. *Natural Language Processing Tools for Reading Level Assessment and Text Simplification for Bilingual Education*. Ph.D. thesis, University of Washington.

Keith Rayner, Kathryn H. Chace, Timothy J. Slattery, and Jane Ashby. 2006. Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, 10(3):241–255.

K. Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124:372–422.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 324–333. Association for Computational Linguistics.

Carlota S. Smith, 1988. *Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered*, chapter Chapter 10: Factors of Linguistic Complexity and Performance, pages 247–279. Lawrence Erlbaum Associates.

W.L. Taylor. 1953. Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30:415–433.

Dustin Tingley, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai. 2014. mediation: R package for causal mediation analysis. *Journal of Statistical Software*, 59(5):1–38.

Katrin Tomanek, Udo Hahn, Steffen Lohmann, and Jürgen Ziegler. 2010. A cognitive cost model of annotations based on eye-tracking data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1158–1167. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2014a. Exploring measures of "readability" for spoken language: Analyzing linguistic features of subtitles to identify age-specific tv programs. In *Proceedings of the Third Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 21–29, Gothenburg, Sweden. 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*, 165(2):142–222.

Sowmya Vajjala. 2015. *Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications*. Ph.D. thesis, University of Tübingen.

Sheila W. Valencia, Karen K. Wixson, and P. David Pearson. 2014. Putting text complexity in context: Refocusing on comprehension of complex text. *The Elementary School Journal*, 115(2):pp. 270–289.

Jacolien van Rij, Martijn Wieling, R. Harald Baayen, and Hedderik van Rijn. 2016. itsadug: Interpreting time series and autocorrelated data using gamms. R package version 2.0.

Sean A. Walmsley, Kathleen M. Scott, and Richard Lehrer. 1981. Effects of document simplification on the reading comprehension of the elderly. In *Journal of Literacy Research*.

Martijn Wieling, Simonetta Montemagni, John Nerbonne, and R Harald Baayen. 2014. Lexical differences between tuscan dialects and standard italian: Accounting for geographic and sociodemographic variation using generalized additive mixed modeling. *Language*, 90(3):669–692.

Gary L. Williamson, Jill Fitzgerald, and A. Jackson Stenner. 2014. Student reading growth illuminates the common core text-complexity standard: Raising both bars. *The Elementary School Journal*, 115(2):pp. 230–254.

S.N Wood. 2006. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

S. N. Wood. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36.