Automatic Readability Assessment
Features, models and their applicability

Sowmya Vajjala

with Detmar Meurers and Julia Hancke

Center for Natural Language Processing
Université Catholique de Louvain, Belgium

28th March 2014

Outline of the Talk

- Background
  - What is readability assessment?
  - Where is it useful?
  - How do we build an automatic model for readability?
- Our Approach
  - Corpus, Features and Modeling
- Applicability to real world
  - for various domains and genres
  - sentences instead of documents
  - practical applications
- Readability Assessment for German

EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School
sfs seminar für sprachwissenschaft

What is readability assessment?

Measuring **how difficult** it is to read a text,

- given a **purpose**, e.g.,
  - identifying appropriate texts for the readers
  - evaluation of natural language generation systems
- based on **properties of the text** using criteria which are
  - theory-driven (e.g., difficult syntactic constructions)
  - data-induced (e.g., corpora with graded texts)
  - in-between (e.g., derived frequency information for words)
- for a **user**, given some information about him (e.g., language ability, age, working memory)
  - obtained directly (e.g., questionnaire), or
  - indirectly (e.g, inferred from nature of a search query)

Where is this useful?

- retrieving relevant reading materials for language learners (e.g., REAP project, http://reap.cs.cmu.edu)
- evaluating the quality of abstracts and automatically generated summaries (Kanungo & Orr 2009)
- readers with special needs (e.g., adults with intellectual disabilities as in Huenerfauth et al. 2009)
- personalized search (Liu et al. 2004)
- assessing difficulty of questions in survey questionnaires (Lenzner 2013)

Where is this useful?

Our work

- text simplification (Vajjala & Meurers, EACL 2014)
  - application context: bilingual classrooms (project starting in June 2014)
- identifying television programs into age-groups based on subtitles (Vajjala & Meurers, PITR 2014)
- comparing the linguistic complexity of German text books across school types & grades (ongoing project)
- analyzing impact of linguistic complexity of questionnaire data on the results (Göllner et al, GEBF 2014 talk)

How is readability assessment done?
Traditional Approaches

- *Traditional readability formulae* used shallow, easy to quantify features
  - e.g., average sentence length, average word length
  - Kincaid et al. (1975), DuBay (2004), ...
- Some others were based on *lexical measures*
  - e.g., words on specific word lists
  - Chall & Dale (1995), ...
- These approaches are popular even now (e.g., in Word).
- Tools exist for calculating scores using
  - popular formulae (e.g., https://readability-score.com)
  - proprietary approaches (e.g., http://www.lexile.com)

How is readability assessment done?

Computational Approaches

- Language n-gram models
  (Si & Callan 2001; Collins-Thompson & Callan 2004; Petersen & Ostendorf 2009)
- Machine learning using lexical and syntactic features
  (Schwarm & Ostendorf 2005; Heilman et al. 2007; Feng 2010)
- Language specific morphological features
  (Dell'Orletta et al. 2011; Francois & Fairon 2012; Hancke et al. 2012)
- Features motivated by cognitive perspective
  (Crossley & McNamara 2011)
- Features motivated by Second Language Acquisition (SLA)
  - measures of language development or proficiency can successfully be used as measures of text complexity
    (Vajjala & Meurers 2012)

EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School
sfs seminar für sprachwissenschaft

Typical Steps in a Machine Learning Approach

- You have some texts annotated with reading level (**corpus**) or you collect some graded texts.
- Start with a hypothesis about text complexity and quantify the hypothesis (**features**).
- How will you build a machine learning model? (**modeling**)
  - Group it in to discrete readability levels (= classification) e.g., beginner, intermediate, advanced
  - continuous readability levels (= regression) e.g., grade 1, 2, . . .
- Verify the performance (**evaluation**) and use this model for predicting the reading level of new texts.

Our Approach: Overview

- **Corpus**: accessible graded corpora (not created by us)
- **Features**: utilize insights on linguistic complexity from Second Language Acquisition and Psycholinguistics
- **Modeling and Evaluation**: build and evaluate regression and classification models
- **Real-life Evaluation**: test applicability by performing multiple evaluations with real-world test sets

The WeeBit Corpus (Vajjala & Meurers 2012)

We combined two reading level annotated corpora to create a new corpus: WeeBit.

Grade Level	*Age in Years*	*Number of Articles*	*Avg. Number of Sentences/Article*
from WeeklyReader			
Level 2	7–8	629	23.41
Level 3	8–9	801	23.28
Level 4	9–10	814	28.12
from BBCBitesize			
KS3	11–14	644	22.71
GCSE	14–16	3500	27.85

Note: If you want to use this corpus for your experiments, please contact me after the talk or send an email later!

EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School
sfs seminar für sprachwissenschaft

Features from Vajjala & Meurers (2012)

- Lexical Features
  - lexical richness features from Second Language Acquisition (SLA) research
    - e.g., Type-Token ratio, noun variation, . . .
  - POS density features
    - e.g., # nouns/# words, # adverbs/# words, . . .
  - traditional features and formulae
    - e.g., # characters per word, Flesch-Kincaid score, . . .
- Syntactic Features
  - syntactic complexity features from SLA research.
    - e.g., # dep. clauses/clause, average clause length, . . .
  - other parse tree features
    - e.g., # NPs per sentence, avg. parse tree height, . . .

Features from Vajjala & Meurers (2014a)

- Morphological properties of words
  - e.g., Does the word contain a stem along with an affix? *abundant=abound+ant*
- Age of Acquisition (AoA)
  - average age-of-acquisition of words in a text
- Other Psycholinguistic features
  - e.g., word abstractness
- Avg. number of senses per word (obtained from WordNet)

Implementation details

Tools, Resources and Algorithms used

- Tools:
  - For Lexical Features
    - Stanford Tagger (Toutanova, Klein, Manning & Singer 2003)
  - For Syntactic Features
    - Berkeley Parser (Petrov & Klein 2007)
    - Tregex Pattern Matcher (Levy & Andrew 2006)
  - For Machine Learning algorithms
    - WEKA (http://www.cs.waikato.ac.nz/ml/weka/)
    - modeled readability as regression, using SMOReg algorithm.
- Resources:
  - Celex Lexical Database (http://celex.mpi.nl)
  - Kuperman et al. (2012)'s AoA ratings
  - MRC Psycholinguistic database (http://ota.oucs.ox.ac.uk/headers/1054.xml)
  - Wordnet Database (http://wordnet.princeton.edu)

How good is our model?

- performance with 10-fold Cross-validation
  - Pearson Correlation = 0.9
  - Room Mean Square Error (RMSE) = 0.53

How good is our model?

- performance with 10-fold Cross-validation
  - Pearson Correlation = 0.9
  - Room Mean Square Error (RMSE) = 0.53
- Performance on a standard test set
  - Common Core Standards example set
  - 168 texts belonging to grade levels 2–12
  - Evaluation: Spearman's rank correlation (scales differ)

How good is our model?

- performance with 10-fold Cross-validation
  - Pearson Correlation = 0.9
  - Room Mean Square Error (RMSE) = 0.53
- Performance on a standard test set
  - Common Core Standards example set
  - 168 texts belonging to grade levels 2–12
  - Evaluation: Spearman's rank correlation (scales differ)

System	**Spearman**
Our System	**0.69**
Nelson et al. (2012):	
REAP (http://reap.cs.cmu.edu)	0.54
ATOS (http://renlearn.com/atos)	0.59
DRP (http://questarai.com/Products/DRPProgram)	0.53
Lexile (http://lexile.com)	0.50
Reading Maturity (http://readingmaturity.com)	**0.69**
SourceRater (http://naeptba.ets.org/SourceRater3)	**0.75**

Does this approach work on real-world texts? (Vajjala & Meurers 2014c)

- How well does our readability model generalize?
- Do our features generalize for other datasets?
- Is there a topic or genre effect?
- What happens if we analyze sentences instead of texts?
  - → How accurately can we identify the differences in the sentential reading level before and after simplification?

Does this approach work on real-world texts?
(Vajjala & Meurers 2014c)

- **How well does our readability model generalize?**
- Do our features generalize for other datasets?
- Is there a topic or genre effect?
- What happens if we analyze sentences instead of texts?

How does our model generalize?
(Vajjala & Meurers 2014c)

- How will we know?
  - test the model with other reading level annotated datasets.

  $\longrightarrow$ created by different groups, with different purposes

How does our model generalize?
(Vajjala & Meurers 2014c)

- How will we know?
  - test the model with other reading level annotated datasets.

  → created by different groups, with different purposes

Test set	Description	Spearman
CommonCore	168 docs, scale: 2-12	0.69
TASA corpus	37K docs, scale: 28-110	0.86
Math Readability Corpus	120 docs, scale: 1-7	0.29

EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School
sfs seminar für sprachwissenschaft

Does this approach work on real-world texts? (Vajjala & Meurers 2014c)

- How well does our readability model generalize?
- **Do our features generalize for other datasets?**
- Is there a topic or genre effect?
- What happens if we analyze sentences instead of texts?

How do the features generalize?
(Vajjala & Meurers 2014c)

- How will we know?
  → We trained regression models with multiple datasets.

How do the features generalize?

(Vajjala & Meurers 2014c)

- How will we know?
  → We trained regression models with multiple datasets.

Corpus	Description	Pearson	RMSE
WeeBit	625 docs per level, 5 levels	0.92	0.53
CommonCore	168 docs, scale: 2-12	0.59	2.69
TASA	37K docs, scale: 28-110	0.97	1.77
Math Readability	120 docs, scale: 1-7	0.51	1.73

Does this approach work on real-world texts? (Vajjala & Meurers 2014c)

- How well does our readability model generalize?
- Do these features generalize for other datasets?
- **Is there a topic or genre effect?**
- What happens if we analyze sentences instead of texts?

Is there a topic or genre effect?
(Vajjala & Meurers 2014c)

Is there a topic or genre effect?
(Vajjala & Meurers 2014c)

- How will we know?
  → Common Core standards test set has genre annotation.

Is there a topic or genre effect? (Vajjala & Meurers 2014c)

- How will we know?
  → Common Core standards test set has genre annotation.
- Genre-wise performance of our model on the dataset:

Genre	Spearman's Correlation
Speech	0.35
Literature	0.51
Informative	0.76
Misc.	0.69

Is there a topic or genre effect? (Vajjala & Meurers 2014c)

- How will we know?
  → Common Core standards test set has genre annotation.
- Genre-wise performance of our model on the dataset:

Genre	Spearman's Correlation
Speech	0.35
Literature	0.51
Informative	0.76
Misc.	0.69

- Can this effect be overcome?

Is there a topic or genre effect?
(Vajjala & Meurers 2014c)

- How will we know?
  - → Common Core standards test set has genre annotation.
- Genre-wise performance of our model on the dataset:

Genre	Spearman's Correlation
Speech	0.35
Literature	0.51
Informative	0.76
Misc.	0.69

- Can this effect be overcome?
  - Yes, by training genre-specific models.
  - → Vajjala & Meurers (2014b)

Tackling the topic or genre effect
(Vajjala & Meurers 2014b)

- We used our feature set to train a model that identifies age-specific TV programs.
- Data: subtitles from 4 BBC channels, that are annotated with three age-groups.
- Result: 96% classification accuracy
- Analysis:
  - single most predictive feature: average AoA of words
  - But accuracy is not reduced if this feature is removed.

  → The classifier is informed by a range of characteristics, not just a single, dominating one.

EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School
sfs seminar für sprachwissenschaft

Does this approach work on real-world texts?

(Vajjala & Meurers 2014c)

- How well does our readability model generalize?
- Do these features generalize for other datasets?
- Is there a topic or genre effect?
- **What happens if we analyze sentences?**

Documents to sentences?

Readability at sentence level

- Does a readability model identify differences between unsimplified and simplified versions of sentences?
- Yes, e.g., consider (sentences from OneStopEnglish.com):

Sentence	**Actual reading level**	**Reading level assigned by our model**
In Beijing, mourners and admirers made their way to lay flowers and light candles at the Apple Store.	*Advanced*	3.6
In Beijing, mourners and admirers came to lay flowers and light candles at the Apple Store.	*Intermediate*	2.3
In Beijing, people went to the Apple Store with flowers and candles.	*Elementary*	1.0

- Why is such sentence level annotation needed?
  - for us: performing and evaluating text simplification

EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School
sfs seminar für sprachwissenschaft

Readability at sentence level

How do we go about this?

- Can our model detect readability at sentence level?
  → Yes, because we do not have any discourse features.
- How do we do that?
  - Approach 1: train a model on sentence level data.
  - Approach 2: use document level model on sentences.
- Approach 1 did not work well. So we explored approach 2.
  - We first studied the distribution of reading levels between sentence pairs.
  - Later used the scores of a document level model to train a sentence ranking model (work in progress).

Readability at sentence level

Two Datasets

- Sentence aligned Wikipedia–Simple Wikipedia dataset, released by Zhu et al. (2010)
- A three level (advanced/intermediate/elementary) sentence aligned corpus.
  - compiled by us from OneStopEnglish.com, consisting of simplified versions of articles from The Guardian

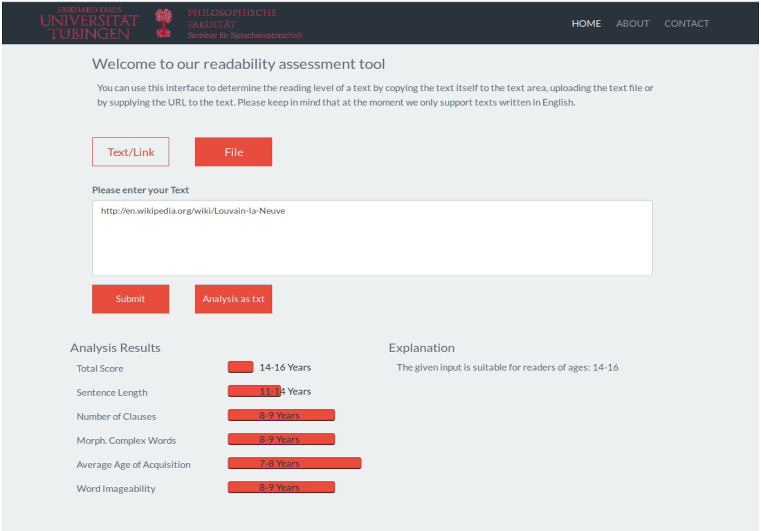
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School
sfs seminar für sprachwissenschaft

Readability at sentence level

What did we learn? (Vajjala & Meurers 2014a)

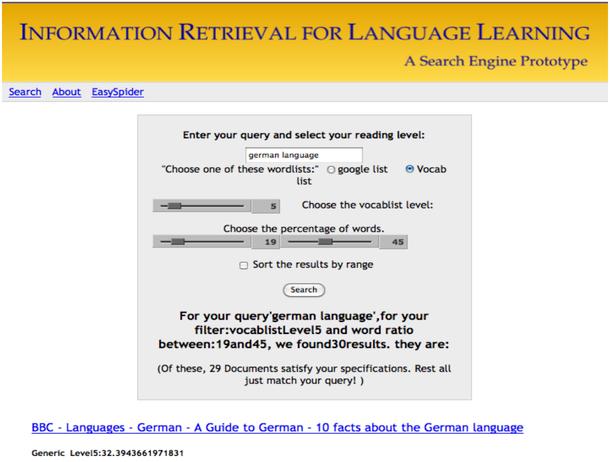
- Simplification is relative
  - → A simplified sentence is simpler than its unsimplified version, but can be harder than another sentence.
- Our approach is particularly successful in classifying the relative reading level of harder sentences.
- It promises to be useful for identifying particularly relevant targets for simplification
  - and to evaluate simplifications given specific readability constraints.

Our Reading level predictor in action



The web-interface is being developed by Ido Friman and will be available soon.

A search engine filtering by vocabulary difficulty



An extension of the Language-Aware Search Engine (Ott & Meurers 2010; Ott 2009).

Readability Assessment beyond English

- Will this approach work for another language?
- Yes, we experimented with German (Hancke et al. 2012).

Readability Assessment for German

Corpus (Hancke, Meurers & Vajjala 2012)

- We created a two class German corpus from two websites containing articles on similar topics:
  - Website for children: GEOlino (http://www.geolino.de)
  - Website for adults: GEO (http://www.geo.de)
  - They are educational monthly magazines covering articles related to technology, nature etc.
  - GEOlino not a simplified version of GEO.
    - Content is created seperately for child readers.
- Modeling: Binary classification.

EBERHARD KARLS UNIVERSITÄT TÜBINGEN
LEAD Graduate School
sfs seminar für sprachwissenschaft

Readability Assessment for German

Features and Model evaluation (Hancke, Meurers & Vajjala 2012)

- Features:
  - **General**: lexical, syntactic, language model features
  - **Language specific**: based on inflectional and derivational morphology of German
- Model performance:
  - Binary classification: 90% accuracy
  - Morphological features alone perform at 85% accuracy.

Current and Future Work

- We are working on combining regression and ranking for determining readability at sentence level,
  - and using this information for automatic text simplification.
- Our related projects in LEAD:
  - Analyzing German school text books across grade levels and school types using readability features
    - with Karin Berendes, Detmar Meurers, Doreen Bryant
  - Studying the cognitive correlates of readability using eye-tracking experiments.
    - with Detmar Meurers, Katharina Scheiter, Alexander Eitel
  - Towards Appropriate Reading Material for Bilingual Classrooms: Evaluating the role of linguistic complexity analysis and text simplification in authentic contexts
    - with Detmar Meurers, Kathrin Jonkmann, Jörg-U. Keßler

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

LEAD Graduate School

sfs seminar für sprachwissenschaft

Thank you!

- Questions? :-)
- Contact: sowmya@sfs.uni-tuebingen.de
- Bibliography: http://purl.org/net/readability-bib.html

References

Chall, J. S. & E. Dale (1995). *Readability Revisted: The New Dale-Chall Readability Formula*. Brookline Books.

Collins-Thompson, K. & J. Callan (2004). A language modeling approach to predicting reading difficulty. In *Proceedings of HLT/NAACL 2004*. Boston, USA. URL http://www.cs.cmu.edu/~callan/Papers/hlt04-kct.pdf.

Crossley, S. A. & D. S. McNamara (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life-Long Learning* 21(2/3), 170–191.

Dell'Orletta, F., S. Montemagni & G. Venturi (2011). READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification. In *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*. pp. 73–83.

DuBay, W. H. (2004). *The Principles of Readability*. Costa Mesa, California: Impact Information. URL http://www.impact-information.com/impactinfo/readability02.pdf.

Feng, L. (2010). Automatic Readability Assessment. Ph.D. thesis, City University of New York (CUNY). URL http://lijun.symptotic.com/files/thesis.pdf?attredirects=0.

Francois, T. & C. Fairon (2012). An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

LEAD Graduate School

sfs seminar für sprachwissenschaft

Hancke, J., D. Meurers & S. Vajjala (2012). Readability Classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. Mumbay, India, pp. 1063–1080. URL http://aclweb.org/anthology-new/C/C12/C12-1065.pdf.

Heilman, M., K. Collins-Thompson, J. Callan & M. Eskenazi (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)*. Rochester, New York, pp. 460–467.

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

Kanungo, T. & D. Orr (2009). Predicting the readability of short web summaries. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. New York, NY, USA: ACM, WSDM '09, pp. 202–211. URL http://doi.acm.org/10.1145/1498759.1498827.

Kincaid, J. P., R. P. J. Fishburne, R. L. Rogers & 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.

Kuperman, V., H. Stadthagen-Gonzalez & M. Brysbaert (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods* 44(4), 978–990. URL http://crr.ugent.be/archives/806.

Lenzner, T. (2013). Are Readability Formulas Valid Tools for Assessing Survey Question Difficulty? *Sociological Methods and Research* -, –.

Levy, R. & G. Andrew (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *5th International Conference on Language Resources and Evaluation*. Genoa, Italy.

Liu, X., W. B. Croft, P. Oh & D. Hart (2004). Automatic recognition of reading levels from user queries. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, SIGIR '04, pp. 548–549. URL http://doi.acm.org/10.1145/1008992.1009114.

Nelson, J., C. Perfetti, D. Liben & M. Liben (2012). *Measures of Text Difficulty: Testing their Predictive Value for Grade Levels and Student Performance*. Tech. rep., The Council of Chief State School Officers. URL http://www.ccsso.org/Documents/2012/Measures20ofText20Difficulty_final.2012.pdf.

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. URL http://drni.de/zap/ma-thesis.

Ott, N. & D. Meurers (2010). Information Retrieval for Education: Making Search Engines Language Aware. *Themes in Science and Technology Education. Special issue on computer-aided language analysis, teaching and learning: Approaches, perspectives and applications* 3(1–2), 9–30. URL http://purl.org/dm/papers/ott-meurers-10.html.

Petersen, S. E. & M. Ostendorf (2009). A machine learning approach to reading level assessment. *Computer Speech and Language* 23, 86–106.

Petrov, S. & D. 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. Rochester, New York, pp. 404–411.

Schwarm, S. & M. Ostendorf (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)*. Ann Arbor, Michigan, pp. 523–530.

Si, L. & J. Callan (2001). A Statistical Model for Scientific Readability. In *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM)*. ACM, pp. 574–576.

Toutanova, K., D. Klein, C. Manning & Y. Singer (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL*. Edmonton, Canada, pp. 252–259.

Vajjala, S. & D. Meurers (2012). On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In J. Tetreault, J. Burstein & C. Leacock (eds.), *In Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*. Montréal, Canada: Association for Computational Linguistics, pp. 163—-173. URL http://aclweb.org/anthology/W12-2019.pdf.

Vajjala, S. & D. 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)*.

Vajjala, S. & D. Meurers (2014b). 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*.

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

LEAD Graduate School

sfs seminar für sprachwissenschaft

Vajjala, S. & D. Meurers (2014c). 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* Thomas François and Delphine Bernhard.

Zhu, Z., D. Bernhard & I. Gurevych (2010). A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of The 23rd International Conference on Computational Linguistics (COLING), August 2010. Beijing, China*.