

ISCL Hauptseminar
Winter Semester 2011

Analyzing complexity and text simplification: Connecting linguistics, processing, and applications

Last update: December 16, 2011

Abstract:

Notions of complexity surface in a number of different contexts: In theoretical linguistics, syntactic structures are analyzed in terms of their complexity and constraints such as the complex-NP constraint are formulated on this basis. In cognitive psychology, the complexity involved in cognitively processing language input in human sentence processing is studied. In second language acquisition research, the analysis of complexity is correlated with stages of acquisition (together with accuracy and fluency). On the applied side, complexity measures have long been used to determine the readability of a given text, and some readability measures have recently been automated in computational linguistics. Relatedly, some proposals for automatic text simplification have been published in recent years, to make information accessible to readers with low reading proficiency.

In this seminar, we will discuss the empirical and conceptual nature of these notions of complexity and explore where the formalization and automatic analysis offered by computational linguistics can lead to applications such as automatic readability measures, search engines supporting the filtering of results by complexity, and automatic text simplification.

Scheduling

Note that the following session plan is subject to change; it only constitutes the current state of our planning as the semester unfolds.

1. Monday, October 17: Organization and Introduction [DETMAR MEURERS]
2. Wednesday, October 19: Overview [SOWMYA V.B.]
3. Monday, October 24: T1 Traditional measures of readability [NIELS OTT]
4. Wednesday, October 26: T1 cont
5. Monday, October 31: T2 psycholinguistics: propositional idea density [LAURA KASSNER]
6. Wednesday, November 2: T3 psycholinguistics: complexity in human sentence processing [DETMAR MEURERS]
7. Monday, November 7: T4
8. Wednesday, November 9: lexical/vocabulary acquisition [SPYRIDOULA GEORGATOU]
9. Monday, November 14:
10. Wednesday, November 16: T7 CAF analysis [STEFANIE WOLF]
11. Monday, November 21: T5 Syntax in L1 acquisition [CHRISTIAN ADAM]

12. Wednesday, November 23: IVAN SAG talk
13. Monday, November 28:
14. Wednesday, November 30: T6 Syntax in L2 acquisition [EDO COLLINS]
15. Monday, December 5: [Tübingen/Berlin Learner Language Workshop](#) (Note special location: [Gästehaus der Universität, Lessingweg 3](#))
16. Wednesday, December 7: T8 discourse features: Coh-Metrix [IULIIA ICHIN-NORBU]
17. Monday, December 12: Machine Learning Background Session [SOWMYA V.B.]
18. Wednesday, December 14: T9 current approaches: REAP ([Brown & Eskenazi 2004](#)) and its approach ([Heilman et al. 2008a](#)) [SOWMYA V.B.]
19. Monday, December 19: T11 current approaches: Peterson/Schwartz/Ostendorf [JULIA HANCKE]
20. Wednesday, December 21:
21. Monday, January 9: T13 current approaches: Feng et al. [SOWMYA V.B.]
22. Wednesday, January 11: Christina Susan Fry (2002): Language complexity, working memory and social intelligence [OLEKSANDR GOZMAN]
23. Monday, January 16: T14 current approaches: DeLite [MICHAEL HAHN]
24. Wednesday, January 18:
25. Monday, January 23: T15 Simplification Intro and early work [KAIDI LOO]
26. Wednesday, January 25: T16 Lexical Simplification [MARYAM GERANMAYEH]
27. Monday, January 30: T17 Syntactic/discourse Simplification [SARAH SCHULZ]
28. Wednesday, February 1:

Instructors:

- Detmar Meurers
 - *Office:* Room 1.28, Blochbau (Wilhelmstr. 19)
 - *Email:* dm@sfs.uni-tuebingen.de
 - *Office hours:* Mondays 11:30–12:30 (best to email beforehand)
- Sowmya V. B.
 - *Office:* Room 1.29, Blochbau (Wilhelmstr. 19)
 - *Email:* sowmya@sfs.uni-tuebingen.de

Course meets: in Seminarraum 1.13, Blochbau (Wilhelmstr. 19)

- Mondays and Wednesdays, 14ct-16

Credits:

- Credit Points: 10 (MA ISCL)

Syllabus (this file):

- [html-Version](#) (<http://purl.org/dm/11/ws/complexity>)

- pdf-Version (<http://purl.org/dm/11/ws/complexity/syllabus.pdf>)

Moodle page: <https://moodle02.zdv.uni-tuebingen.de/course/view.php?id=96>

Nature of course and our expectations: This is a Hauptseminar which on the one hand intends to provide an overview of current perspectives and approaches on complexity in linguistics, psycholinguistics, and computational linguistics. On the other hand, the computational linguistics students enrolled in the course are expected to define and implement an approach for complexity analysis or text simplification as their term paper project.

1. regularly and actively participate in class, read the papers assigned by any of the presenters and post a question on Moodle to the “Reading Discussion Forum” on each reading *at the latest on the day before it is discussed* in class. (30% of grade for Hauptseminar)

Note: Following the general university rules, missing more than two meetings unexcused, automatically results in failing the class.

2. explore and present a topic (30% of grade for Hauptseminar):

- select one of the sub-topics during the first week of the semester
- thoroughly research the topic, taking our literature pointers *as a starting point*
- prepare the presentation with slides and discuss the presentation with the instructor during the office hours, generally *a week before the presentation*
- start a new Moodle thread on the “Reading Discussion Forum” specifying what every course participant should read to prepare for your presentation *a week before your presentation*
- present the topic in class

3. work out a project term paper (40% of grade)

- select a topic and submit a one-page abstract *by January 25, 2012*
 - For computational linguistics students, the topic of the paper in general will be the exploration and implementation of an approach analyzing the complexity or performing text simplification.
- email the term paper in pdf format to the instructor *before the beginning of the next semester, i.e., by March 30, 2012.*
 - Note for ISCL students: The term paper must be produced in LaTeX, and BibTex must be used for the bibliography.

Academic conduct and misconduct: Research is driven by discussion and free exchange of ideas, motivations, and perspectives. So you are encouraged to work in groups, discuss, and exchange ideas. At the same time, the foundation of the free exchange of ideas is that everyone is open about where they obtained which information. Concretely,

this means you are expected to always make explicit when you've worked on something as a team – and keep in mind that being part of a team always means sharing the work.

For text you write, you always have to provide explicit references for any ideas or passages you reuse from somewhere else. Note that this includes text “found” on the web, where you should cite the url of the web site in case no more official publication is available.

Class etiquette: Please do not read or work on materials for other classes in our seminar. Come to class on time and do not pack up early. When our seminar meets in the computer lab, only use the computers when you are asked to do a specific activity – do not read email or browse the web. All portable electronic devices such as cell phones should be switched off for the entire length of the flight, oops, class. If for some reason, you must leave early or you have an important call coming in, or you have to miss class for an important reason, please let the instructor know *before* class.

Topics:

I. Complexity/Difficulty

- Research strands:
 - **T1** Traditional readability measures (Flesch 1948; Dale & Chall 1948; Coleman & Liau 1975; Kincaid et al. 1975; DuBay 2004, 2006; Bennöhr 2007)
Zipf (1936): longer words are less frequent
 - Psycholinguistic measures of complexity
 - * interesting early work (Monkhouse 1972)
 - * **T2** propositional idea density (Brown et al. 2008), cf. also syntactic information density (Jaeger 2010)
 - * **T3** sentence comprehension difficulty in human sentence processing (Boston et al. 2008, 2011)
 - dissociation of word reading and text comprehension (Oakhill et al. 2003)
 - (?)
 - Language complexity, working memory and social intelligence (University of Newcastle upon Tyne, UK; 2002) by Christina Susan Fry
<http://www.hedweb.com/bgcharlton/tina-fry.html>
 - Measures of language acquisition
 - * **T4** lexical/vocabulary acquisition:
 - (Milton 2009) book on second language vocabulary acquisition
 - predicting level of learners using word lists (Pendar & Chapelle 2008), cf. also: (Chall & Dale 1995; Coxhead 2000; West 1953; Hancioglu et al. 2008; Cobb 2010),
 - Lexical Diversity Measures (McCarthy 2005; McCarthy & Jarvis 2010), cf. also CELEX database (Baayen et al. 1995), and for German the DWDS <http://www.dwds.de/> and <http://dlexdb.de/> as well as Tschirner (2008)

- measuring active vocabulary of learners: Lexical Frequency Profiles ([Laufer & Nation 1995](#)), cf. also: *Range* tool by Paul Nation: <http://www.victoria.ac.nz/lals/staff/paul-nation.aspx>
- * Syntax:
 - **T5** Child language acquisition: Revised D-Level ([Lu 2009](#); [Voss 2005](#)), IPSYN ([Sagae et al. 2005](#))
 - **T6** second language learners ([?Lu 2010, 2011](#); [Schulze 2010](#); [Schulze et al. 2010](#); [Vyatkina 2012](#))
 - **T7** Complexity in CAF (Complexity, Accuracy and Fluency) analysis of learner language ([Wolfe-Quintero et al. 1998](#); [Ortega 2003](#); [Housen & Kuiken 2009](#))
- * **T8** discourse:
 - Coh-Metrix Project (<http://cohmetrix.memphis.edu>, [McNamara et al. 2002](#); [Crossley et al. 2000, 2008](#))
- Current computational linguistic approaches
 - **T9** REAP ([Heilman et al. 2008b](#); [Brown & Eskenazi 2004, 2005](#); [Collins-Thompson & Callan 2004, 2005](#); [Si & Callan 2001](#); [Heilman et al. 2007, 2008a](#); [Dela Rosa & Eskenazi 2011](#))
 - **T10** ETS SourceFinder ([Sheehan et al. 2007, 2008, 2009, 2010](#))
 - **T11** ([Schwartz & Ostendorf 2005](#); [Petersen & Ostendorf 2006a,b](#); [Petersen 2007](#); [Petersen & Ostendorf 2007, 2009](#))
 - **T13** ([Feng 2010](#); [Feng et al. 2009](#); [Lijun Feng & Elhadad 2010](#); [Pitler & Nenkova 2008](#))
 - **T14** an approach targeting German ([vor der Brück et al. 2008](#); [Vor der Brück 2008](#))
- some other issues/projects
 - Read-X ([Miltsakaki & Troutt 2007, 2008](#); [Miltsakaki 2009](#))
 - searching texts suitable for language learners ([Sharoff et al. 2008](#); [Ozasa et al. 2007, 2008](#); [Newbold et al. 2010](#); [Ott & Meurers 2010](#); [Bendersky et al. 2011](#))
 - for special needs users ([Kanungo & Orr 2009](#); [Huenerfauth et al. 2009](#))
 - for text simplification ([Aluisio et al. 2007](#); [Jonnalagadda et al. 2009](#); [Crossley et al. 2011](#))
 - **T19** user modeling ([Liu et al. 2004](#); [Nakatani et al. 2009, 2010](#)), Michael Welmsley's work on vocabulary learning adapted to learner
 - languages other than English: Chinese ([Lau 2006](#)), Japanese ([Sato et al. 2008](#)), Dutch ([van Oosten et al. 2010](#))
 - evaluation
 - * Evaluation techniques used in statistical approaches
 - * Comparison and correlation between and against various traditional measures ([van Oosten et al. 2010](#); [Van Oosten et al. 2011](#))

- * Cloze test based evaluation
- corpora used
 - * BBC Bitesize (<http://www.bbc.co.uk/schools/bitesize>)
 - * educational classroom magazines (e.g., Weekly Reader, <http://www.weeklyreader.com>, [onestopenglish.com](http://www.onestopenglish.com), [readinga-z.com](http://www.readinga-z.com))
 - * DeLite Corpus (vor der Brück et al. 2008; Vor der Brück 2008)
 - * (Miller & Coleman 1967) describes the preparation of a corpora of 36 prose passages of 150 words each, prepared by using cloze technique

II. Simplification

- **T15** Introduction and early work
 - Feng (2008) provides a detailed summary of pre-2008 work of all sorts
 - (Chandrasekar et al. 1996; R.Chandrasekar 1996)
 - An architecture for a simplification system (Siddharthan 2002, 2004)
- **T16** Lexical Simplification Approaches (Jan De Belder 2010; Yatskar et al. 2010; Biran et al. 2011)
- **T17** Approaches to syntactic (Jonnalagadda et al. 2009; R.Chandrasekar 1996; Klebanov et al. 2004) and discourse/text structure simplification (Carroll et al. 1998; Canning 2002; Inui et al. 2003; Devlin & Unthank 2006; Williams & Reiter 2008)
- **T18** Applications (Canning & Tait 1999; Klebanov et al. 2004; Bouayad-Agha et al. 2006; Aluisio et al. 2007; Burstein et al. 2007; Gasperin et al. 2009b,a)
- On building corpora (Petersen & Ostendorf 2007; S & H 2011; Coster & Kauchak 2011)
- Evaluation and Corpora (Jonnalagadda & Gonzalez 2009; Yatskar et al. 2010; Cohn et al. 2005)
- Some common corpora used in these works so far are:
 - English Wiki-Simple Wiki
 - Weekly Reader corpus
 - KidsPost-WashingtonPost
 - Enc.Britannica-Enc.Elementary
 - Time-TimeForKids

References

- Aluisio, S., L. Specia, C. Gasperin & C. Scarton (2007). Readability Assessment for Text Simplification.

- Baayen, R. H., R. Piepenbrock & L. Gulikers (1995). The CELEX Lexical Database (CD-ROM). CDROM. URL http://www.ldc.upenn.edu/Catalog/readme_files/celex.readme.html.
- Bendersky, M., W. B. Croft & Y. Diao (2011). Quality-biased ranking of web documents. In *Proceedings of the fourth ACM international conference on Web search and data mining*. New York, NY, USA: ACM, WSDM '11, pp. 95–104. URL <http://doi.acm.org/10.1145/1935826.1935849>.
- Bennöhr, J. (2007). A Web-Based Personalised Textfinder for Language Learners. In G. Rehm, A. Witt & L. Lemnitzer (eds.), *Data Structures for Linguistic Resources and Applications*. Tübingen: Gunter Narr Verlag.
- Biran, O., S. Brody & N. Elhadad (2011). Putting it Simply: a Context-Aware Approach to Lexical Simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 496–501. URL <http://www.aclweb.org/anthology/P11-2087>.
- Boston, M. F., J. T. Hale, U. Patil, R. Kliegl & S. Vasishth (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research* 2(1), 1–12. URL <http://www.jemr.org/online/2/1/1>.
- Boston, M. F., J. T. Hale, S. Vasishth & R. Kliegl (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes* 26(3), 301–349.
- Bouayad-Agha, N., A. Gil, O. Valentin & V. Pascual (2006). A Sentence Compression Module for Machine-Assisted Subtitling. In *CICLING*. pp. 490–501.
- Brown, C., T. Snodgrass, S. J. Kemper, R. Herman & M. A. Covington (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods* 40(2), 540–545.
- Brown, J. & M. Eskenazi (2004). Retrieval of authentic documents for reader-specific lexical practice. In R. Delmonte (ed.), *InSTIL/ICALL 2004 Symposium on Computer Assisted Learning, NLP and speech technologies in advanced language learning systems*. Venice, Italy: International Speech Communication Association (ISCA). URL <http://reap.cs.cmu.edu/Papers/InSTIL04-jonbrown.pdf>.
- Brown, J. & M. Eskenazi (2005). Student, text and curriculum modeling for reader-specific document retrieval. In *Proceedings of the IASTED International Conference on Human-Computer Interaction*. Phoenix, Arizona. URL http://www.cs.cmu.edu/~max/mainpage_files/2005-REAP-IASTED-HCI.pdf.
- Burstein, J., J. Shore, J. Sabatini, Y.-W. Lee & M. Ventura (2007). The Automated Text Adaptation Tool. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. Association for Computational Linguistics, pp. 3–4. URL <http://www.aclweb.org/anthology-new/N/N07/N07-4002.pdf>.
- Canning, Y. (2002). Syntactic Simplification of Text. Ph.D. thesis, University of Sunderland.
- Canning, Y. & J. Tait (1999). Syntactic Simplification of Newspaper Text for Aphasic Readers. In *Proceedings of SIGIR-99 Workshop on Customised Information Delivery*. Berkeley, CA, pp. 6–11.

- Carroll, J., G. Minnen, Y. Canning, S. Devlin & J. Tait (1998). Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*. Madison, Wisconsin: Association for the Advancement of Artificial Intelligence (AAAI). URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.51.1145&rep=rep1&type=pdf>.
- Chall, J. S. & E. Dale (1995). *Readability Revisted: The New Dale-Chall Readability Formula*. Brookline Books.
- Chandrasekar, R., C. Doran, B. Srinivas & R. Ch (1996). Motivations and Methods for Text Simplification. URL <http://acl.ldc.upenn.edu/C/C96/C96-2183.pdf>.
- Cobb, T. (2010). Learning about language and learners from computer programs. *Reading in a Foreign Language* 22(1), 181–200. URL <http://nflrc.hawaii.edu/RFL/April2010/articles/cobb.pdf>.
- Cohn, T., C. Callison-burch & M. Lapata (2005). Constructing Corpora for the Development and Evaluation of Paraphrase Systems.
- Coleman, M. & T. Liau (1975). A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology* 60, 283–284.
- 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>.
- Collins-Thompson, K. & J. Callan (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology* 56(13), 1448–1462.
- Coster, W. & D. Kauchak (2011). Simple English Wikipedia: A New Text Simplification Task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 665–669. URL <http://www.aclweb.org/anthology/P11-2117.pdf>.
- Coxhead, A. (2000). A New Academic Word List. *Teachers of English to Speakers of Other Languages* 34(2), 213–238. URL <http://www.ingentaconnect.com/content/tesol/tq/2000/00000034/00000002/art00002?token=004112cb2e405847447b492b2f5f737b6f2c2b67217d3375686f236efca0b0d66>.
- Crossley, S. A., D. B. Allen & D. McNamara (2011). Text Readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language* 23(1), pp. 84–101.
- Crossley, S. A., D. F. Dufty, P. M. McCarthy & D. S. McNamara (2000). Toward a New Readability : A Mixed Model Approach. In S. McNamara & G. Trafton (eds.), *Proceedings of the 29th annual conference of the Cognitive Science Society. Austin, TX*. Cognitive Science Society, pp. 197–202.
- Crossley, S. A., J. Greenfield & D. S. McNamara (2008). *Assessing Text Readability Using Cognitively Based Indices*, Teachers of English to Speakers of Other Languages, Inc. 700 South Washington Street Suite 200, Alexandria, VA 22314, pp. 475–493.
- Dale, E. & J. S. Chall (1948). A Formula for Predicting Readability. *Educational research bulletin; organ of the College of Education* 27(1), 11–28.
- Dela Rosa, K. & M. Eskenazi (2011). Effect of Word Complexity on L2 Vocabulary

- Learning. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*. Portland, Oregon: Association for Computational Linguistics, pp. 76–80. URL <http://www.aclweb.org/anthology/W11-1409>.
- Devlin, S. & G. Unthank (2006). Helping aphasic people process online information. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*. New York, NY, USA: ACM, Assets '06, pp. 225–226. URL <http://doi.acm.org/10.1145/1168987.1169027>.
- DuBay, W. H. (2004). *The Principles of Readability*. Costa Mesa, California: Impact Information. URL <http://www.impact-information.com/impactinfo/readability02.pdf>.
- DuBay, W. H. (2006). *The Classic Readability Studies*.
- Feng, L. (2008). *Text Simplification: A Survey*. Tech. rep., CUNY. URL <http://lijun-sympotic.com/files/TextSimplification.pdf>.
- Feng, L. (2010). Automatic Readability Assessment. Ph.D. thesis, CUNY.
- Feng, L., N. Elhadad & M. Huenerfauth (2009). Cognitively Motivated Features for Readability Assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. Athens, Greece: Association for Computational Linguistics, pp. 229–237. URL <http://www.aclweb.org/anthology/E09-1027>.
- Flesch, R. F. (1948). A New Readability Yardstick. *Journal of Applied Psychology* 32(3), 221–233.
- Gasperin, C., E. Maziero, L. Specia, P. T.S.P. & S. Aluisio (2009a). Natural language processing for social inclusion: a text simplification architecture for different literacy levels. In *XXXVI Seminário Integrado de Software e Hardware (SEMISH-2009)*. Bento Gonçalves, Brazil, pp. 387–401.
- Gasperin, C., L. Specia, T. F. Pereira & S. M. Aluisio (2009b). Learning When to Simplify Sentences for Natural Text Simplification. In *Encontro Nacional de Inteligência Artificial (ENIA-2009)*.
- Hancioglu, N., S. Neufeld & J. Eldridge (2008). Through the looking glass and into the land of lexico-grammar. *English for Specific Purposes* 27(4), 459 – 479. URL <http://www.sciencedirect.com/science/article/pii/S0889490608000355>.
- 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: Association for Computational Linguistics, pp. 460–467. URL <http://aclweb.org/anthology/N07-1058>.
- Heilman, M., K. Collins-Thompson & M. Eskenazi (2008a). An Analysis of Statistical Models and Features for Reading Difficulty Prediction. In *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications*. Columbus, Ohio. URL <http://aclweb.org/anthology/W08-0909>.
- Heilman, M., L. Zhao, J. Pino & M. Eskenazi (2008b). Retrieval of Reading Materials for Vocabulary and Reading Practice. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08*. Columbus, Ohio: Association for Computational Linguistics, pp. 80–88. URL <http://aclweb.org/anthology/W08-0910>.

- Housen, A. & F. Kuiken (2009). Complexity, Accuracy, and Fluency in Second Language Acquisition. *Applied Linguistics* 30(4), 461–473. URL <http://applij.oxfordjournals.org/content/30/4/461.full.pdf>.
- 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>.
- Inui, K., A. Fujita, T. Takahashi, R. Iida & T. Iwakura (2003). Text Simplification for Reading Assistance: A Project Note. In *Proceedings of the Second International Workshop on Paraphrasing, held at ACL 2003*. URL <http://aclweb.org/anthology/W03-1602>.
- Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology* 61(1), 23–62.
- Jan De Belder, M.-F. M. (2010). Text Simplification For Children. In *SIGIR workshop on accessible search systems, 2010*.
- Jonnalagadda, S. & G. Gonzalez (2009). Sentence Simplification Aids Protein-Protein Interaction Extraction. In *Proceedings of The 3rd International Symposium on Languages in Biology and Medicine, Jeju Island, South Korea, November 8-10, 2009*.
- Jonnalagadda, S., L. Tari, J. Hakenberg, C. Baral & G. Gonzalez (2009). Towards Effective Sentence Simplification for Automatic Processing of Biomedical Text. In *Proceedings of the NAACL-HLT 2009, Boulder, USA, June*.
- 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.
- Klebanov, B. B., K. Knight & D. Marcu (2004). Text Simplification for Information-Seeking Applications. In *On the Move to Meaningful Internet Systems, Lecture Notes in Computer Science*. Springer Verlag, pp. 735–747. URL <http://www.isi.edu/~marcu/papers/factoids04.pdf>.
- Lau, T. P. (2006). Chinese Readability Analysis and Its Applications on the Internet. Master's thesis, CUHK, Hongkong.
- Laufer, B. & P. Nation (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics* 16(3), 307–322. URL <http://applij.oxfordjournals.org/content/16/3/307.abstract>.
- Lijun Feng, Martin Jansche, M. H. & N. Elhadad (2010). A Comparison of Features for Automatic Readability Assessment. In *In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China..* URL <http://www.aclweb.org/anthology/C10-2032>.
- 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>.
- Lu, X. (2009). Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics* 14(1), 3–28.
- Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15(4), 474–496.
- Lu, X. (2011). A Corpus-Based Evaluation of Syntactic Complexity Measures as Indices of College-Level ESL Writers' Language Development. *TESOL Quarterly* 45(1), 36–62.
- McCarthy, P. & S. Jarvis (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 42(2), 381–392.
- McCarthy, P. M. (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). Ph.D. thesis, University of Memphis. URL <https://umdrive.memphis.edu/pmmccrth/public/Papers/MTLD%20dissertation.doc>.
- McNamara, D. S., M. M. Louwerse & A. C. Graesser (2002). Coh-Metrix: Automated Cohesion and Coherence Scores to Predict Text Readability and Facilitate Comprehension. Proposal of Project funded by the Office of Educational Research and Improvement, Reading Program. URL <http://cohmetrix.memphis.edu/cohmetrixpr/archive/Coh-MetrixGrant.pdf>.
- Miller, G. & E. Coleman (1967). A SET OF THIRTY-SIX PROSE PASSAGES CALIBRATED FOR COMPLEXITY. *Journal of Verbal Learning and Verbal Behavior* 6(6), 851–854.
- Milton, J. (2009). *Measuring Second Language Vocabulary Acquisition*. SLA. Multilingual Matters Ltd.
- Miltsakaki, E. (2009). Matching readers' preferences and reading skills with appropriate web texts. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*. Stroudsburg, PA, USA: Association for Computational Linguistics, EACL '09, pp. 49–52. URL <http://dl.acm.org/citation.cfm?id=1609049.1609062>.
- Miltsakaki, E. & A. Troutt (2007). Read-X: Automatic Evaluation of Reading Difficulty of Web Text. In T. Bastiaens & S. Carliner (eds.), *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2007*. Quebec City, Canada: AACE, pp. 7280–7286. URL <http://www.editlib.org/p/26932>.
- Miltsakaki, E. & A. Troutt (2008). Real Time Web Text Classification and Analysis of Reading Difficulty. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications (BEA-3) at ACL'08*. Columbus, Ohio: Association for Computational Linguistics, pp. 89–97. URL <http://aclweb.org/anthology/W08-0911>.
- Monkhouse, K. M. (1972). Measures of Syntactic Complexity and Their Relationship to Psychological Scale Values of Systematically Varied Sentences. Ph.D. thesis, The University of Iowa, Speech Pathology.
- Nakatani, M., A. Jatowt & K. Tanaka (2009). Easiest-First Search: Towards Comprehension-based Web Search. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*. ACM Press, Hong Kong, China,

- pp. 2057–2060. URL <http://www.dl.kuis.kyoto-u.ac.jp/~adam/cikm09.pdf>.
- Nakatani, M., A. Jatowt & K. Tanaka (2010). Adaptive Ranking of Search Results by Considering User's Comprehension. In *Proceedings of the 4th International Conference on Ubiquitous Information Management and Communication (ICUIMC 2010)*. ACM Press, Suwon, Korea, pp. 182–192. URL <http://www.dl.kuis.kyoto-u.ac.jp/~adam/icuimc10.pdf>.
- Newbold, N., H. McLaughlin & L. Gillam (2010). Rank by Readability: Document Weighting for Information Retrieval. In H. Cunningham, A. Hanbury & S. Rüger (eds.), *Advances in Multidisciplinary Retrieval*, Springer Berlin / Heidelberg, vol. 6107 of *Lecture Notes in Computer Science*, pp. 20–30. URL http://dx.doi.org/10.1007/978-3-642-13084-7_3.
- Oakhill, J., K. Cain & P. Bryant (2003). The dissociation of word reading and text comprehension: Evidence from component skills. *Language and Cognitive Processes* 18(4), 443–468. URL <http://www.tandfonline.com/doi/abs/10.1080/01690960344000008>.
- Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics* 24(4), 492–518.
- 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>.
- Ozasa, T., G. Weir & M. Fukui (2008). Toward a Readability Index for Japanese Learners of EFL. In *Proceedings of the 13th Conference of Pan-Pacific Association of Applied Linguistics (PAAL'08)*. University of Hawaii, Manoa: Pan-Pacific Association of Applied Linguistics. URL http://www.cis.strath.ac.uk/cis/research/publications/papers/strath_cis_publication_2263.pdf.
- Ozasa, T., G. R. S. Weir & M. Fukui (2007). Measuring Readability for Japanese Learners of English. In *Proceedings of the 12th Conference of Pan-Pacific Association of Applied Linguistics*. Pan-Pacific Association of Applied Linguistics. URL <http://www.paaljapan.org/conference2007/index.html>.
- Pendar, N. & C. Chapelle (2008). Investigating the Promise of Learner Corpora: Methodological Issues. *CALICO Journal* 25(2), 189–206. URL https://calico.org/html/article_689.pdf.
- Petersen, S. E. (2007). Natural Language Processing Tools for Reading Level Assessment and Text Simplification for Bilingual Education. Ph.D. thesis, University of Washington. URL http://sarahpetersen.net/sarah_petersen_dissertation.pdf.
- Petersen, S. E. & M. Ostendorf (2006a). Assessing the Reading Level of Web Pages. In *Ninth International Conference on Spoken Language Processing (Interspeech-ICSLP)*. Pittsburgh, Pennsylvania. URL http://sarahpetersen.net/portfolio/petersen_is2006.pdf.
- Petersen, S. E. & M. Ostendorf (2006b). Assessing the Reading Level of Web Pages (Poster). In *Proceedings of Interspeech 2006*.
- Petersen, S. E. & M. Ostendorf (2007). Text Simplification for Language Learners: A Corpus Analysis. In *Speech and Language Technology for Education (SLaTE)*. URL http://sarahpetersen.net/portfolio/Petersen_Ostendorf_SLaTE2007_final.pdf.
- Petersen, S. E. & M. Ostendorf (2009). A machine learning approach to reading level

- assessment. *Computer Speech and Language* 23, 86–106.
- Pitler, E. & A. Nenkova (2008). Revisiting readability: a unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, EMNLP '08, pp. 186–195. URL <http://dl.acm.org/citation.cfm?id=1613715.1613742>.
- R.Chandrasekar, B. (1996). *Automatic Induction of Rules for Text Simplification*. Tech. Rep. IRCS Report 96–30, Upenn, NSF Science and Technology Center for Research in Cognitive Science.
- S, B. & S. H (2011). An Unsupervised Alignment Algorithm for Text Simplification Corpus Construction. In *ACL Workshop on Monolingual Text-to-Text Generation*.
- Sagae, K., A. Lavie & B. MacWhinney (2005). Automatic measurement of syntactic development in child language. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'05)*. Ann Arbor, MI. URL <http://aclweb.org/anthology/P05-1025>.
- Sato, S., S. Matsuyoshi & Y. Kondoh (2008). Automatic Assessment of Japanese Text Readability Based on a Textbook Corpus. In *LREC'08*. pp. –1–1. URL www.lrec-conf.org/proceedings/lrec2008/pdf/165_paper.pdf.
- Schulze, M. (2010). Measuring textual complexity in student writing. Talk slides for a presentation at the 2010 conference of the American Association for Applied Linguistics (AAAL), Atlanta, GA. URL http://wcgs.ca/~mschulze/papers/aaal_2010.pdf.
- Schulze, M., P. Wood & B. Pokorny (2010). Measuring balanced complexity. unpublished manuscript.
- Schwartz, 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. URL <http://aclweb.org/anthology/P05-1065>.
- Sharoff, S., S. Kurella & A. Hartley (2008). Seeking needles in the Web's haystack: Finding texts suitable for language learners. In *Proceedings of the 8th Teaching and Language Corpora Conference (TaLC-8)*. Lisbon, Portugal.
- Sheehan, K. M., I. Kostin & Y. Futagi (2008). When Do Standard Approaches for Measuring Vocabulary Difficulty, Syntactic Complexity and Referential Cohesion Yield Biased Estimates of Text Difficulty? In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*. URL <http://csjarchive.cogsci.rpi.edu/proceedings/2008/pdfs/p1978.pdf>.
- Sheehan, K. M., I. Kostin & Y. Futagi (2009). When Do Standard Approaches for Measuring Vocabulary Difficulty, Syntactic Complexity and Referential Cohesion Yield Biased Estimates of Text Difficulty? In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*. URL <http://csjarchive.cogsci.rpi.edu/proceedings/2008/pdfs/p1978.pdf>.
- Sheehan, K. M., I. Kostin, Y. Futagi & M. Flor (2010). *Generating Automated Text Complexity Classifications That Are Aligned with Targeted Text Complexity Standards*. Tech. Rep. RR-10-28, ETS. URL http://www.ets.org/research/policy_research_reports/rr-10-28.
- Sheehan, K. M., I. W. Kostin & Y. Futagi (2007). SourceFinder: A Construct-Driven Approach for Locating Appropriately Targeted Reading Comprehension Source

- Texts. In *Proceedings of the 2007 Workshop of the International Speech Communication Association, Special Interest Group on Speech and Language Technology in Education*. URL http://www.eee.bham.ac.uk/russellm/SLaTE2007/SLaTE07_Sheehan_SourceFinder.pdf.
- 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. Poster description.
- Siddharthan, A. (2002). An Architecture for a Text Simplification System. In *In Proceedings of the Language Engineering Conference 2002 (LEC 2002)*.
- Siddharthan, A. (2004). *Syntactic simplification and text cohesion*. Tech. Rep. UCAM-CL-TR-597, University of Cambridge Computer Laboratory. URL <http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-597.pdf>.
- Tschirner, E. (2008). Das professionelle Wortschatzminimum im Deutschen als Fremdsprache. *Deutsch als Fremdsprache* 45, 195–208.
- Van Oosten, P., V. Hoste & D. Tanghe (2011). A posteriori agreement as a quality measure for readability prediction systems. In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part II*. Berlin, Heidelberg: Springer-Verlag, CICLing'11, pp. 424–435. URL <http://dl.acm.org/citation.cfm?id=1964750.1964790>.
- van Oosten, P., D. Tanghe & V. Hoste (2010). Towards an Improved Methodology for Automated Readability Prediction. In *LREC'10*. pp. –1–1. URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/286_Paper.pdf.
- vor der Brück, T., S. Hartrumpf & H. Helbig (2008). A Readability Checker with Supervised Learning using Deep Syntactic and Semantic Indicators.
- Vor der Brück, T. H. H. J. L. (2008). *The readability checker DeLite*. Tech. Rep. Technical Report 345-5/2008, Fakultät für Mathematik und Informatik, FernUniversität in Hagen.
- Voss, M. J. (2005). *Determining Syntactic Complexity Using Very Shallow Parsing*. Research Report 2005-01, Computer Analysis of Speech for Psychological Research (CASPR), Institute for Artificial Intelligence, The University of Georgia. URL <http://www.ai.uga.edu/caspr/2005-01-Voss.pdf>. Published verison of MSc thesis.
- Vyatkina, N. (2012). The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *The Modern Language Journal* To appear.
- West, M. (1953). *A General Service List of English Words*. London: Longmans.
- Williams, S. & E. Reiter (2008). Generating basic skills reports for low-skilled readers*. *Nat. Lang. Eng.* 14, 495–525. URL <http://dl.acm.org/citation.cfm?id=1520025.1520029>.
- Wolfe-Quintero, K., S. Inagaki & H.-Y. Kim (1998). *Second Language Development in Writing: Measures of Fluency, Accuracy & Complexity*. Honolulu: Second Language Teaching & Curriculum Center, University of Hawaii at Manoa.
- Yatskar, M., B. Pang, C. Danescu-Niculescu-Mizil & L. Lee (2010). For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Proceedings of the NAACL*. pp. 365–368.
- Zipf, G. K. (1936). *The Psycho-Biology of Language*. London: Routledge.