Core CL Hauptseminar **Winter Semester 2016**

Computational Linguistic Analysis of Linguistic Complexity

Last update: December 18, 2016

Abstract:

Aspects of complexity are important under a number of different theoretical and applied perspectives related to language - from theoretical linguistics making reference to complex noun phrases and recursion, via language acquisition research discussing complexity as a measure of development, or readability research distinguishing which audience a text is appropriate for and how it could be simplified, to psycholinguistic research on human sentence processing computing surprisal and other measures reflecting processing difficulty. Interestingly, complexity is an issue at all levels of linguistic modeling, including the lexicon and morphology, syntax, semantics, and discourse as well as aspects of language use such as frequency. In this Hauptseminar, we will investigate and develop computational linguistic techniques and applications supporting the automatic identification of a broad range of aspects of linguistic complexity, including computational models of human processing and modules needed to build tools for readability classification, simplification, or information retrieval.

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. Wednesday, October 26: Organization and Overview [Detmar Meurers]
2. Friday, October 28: Introduction [Detmar Meurers]
3. Wednesday, November 2: Introduction [Detmar Meurers]
4. Friday, November 4: *no class*
5. Wednesday, November 9: Introduction [Detmar Meurers]
6. Friday, November 11: Traditional readability measures [Ekaterina Panfilova]
   - (DuBay 2004, 2006; François & Miltsakaki 2012)
7. Wednesday, November 16: Psycholinguistic Measures
   - Eye tracking background [Zarah Solgi]
8. Friday, November 18: Psycholinguistic Measures II:
   - Dependency Locality Theory [Matthias Karlbaer]
9. Wednesday, November 23: Psycholinguistic Measures III:
   - Surprisal
     - (Boston et al. 2008) [Lukas Walter]
     - (Boston et al. 2011) [Kuan Yu]

10. Friday, November 25: Psychological Models of Comprehension
    - Kintsch's Construction Integration model of reading (Kintsch & van Dijk 1978; Kintsch 1988) [Nora Stefanova Kumpikova]
    - Propositional Idea Density (Brown et al. 2008) [Tobias Elssner]
11. Wednesday, November 30: SLA Background on CAF: Complexity, Accuracy, and Fluency
    - (Skehan 1989) [Miriam Marthaler]
12. Friday, December 2: SLA Background on CAF: Complexity, Accuracy, and Fluency
    - (Wolfe-Quintero et al. 1998; Housen & Kuiken 2009) [Sarah Taylor]
13. Wednesday, December 7: CAF
    - (Ortega 2003) [Jonas Schäfer]
14. Friday, December 9: Lexical measures in SLA
    - (Kyle & Crossley 2015) [Ekaterina Lazaruk]
    - (Lu 2012) [Mei-Shin Wu]
    - some related work: (Laufer & Nation 1995; Malvern et al. 2004; McCarthy & Jarvis 2010; Read & Nation 2004)
15. Wednesday, December 14: Syntactic complexity in SLA
    - (Covington et al. 2006; Lu 2010) [Teslin Roys]
    - (Cheung & Kemper 1992) [Jonas Raggatz]
16. Friday, December 16: Discourse and CohMetrix
    - (McNamara et al. 2002; Graesser et al. 2004) [Holger Muth-Hellebrandt]
    - (Crossley et al. 2000, 2008) [Alexander Hartmann]
17. Wednesday, December 21: Discourse and CohMetrix II
    - Connectives [Ryan Callihan]
18. Wednesday, January 11: Analysis and Task effects
    - (Vyatkina 2012) [Frank Obeng]
    - (Alexopoulou et al. submitted) [Rebecca Long]
19. Friday, January: 13: ETS SourceFinder (Sheehan et al. 2007, 2008, 2009, 2010) [Andreas Daul]
20. Wednesday, January 18: 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) [Sarah Schneider]
21. Friday, January: 20: German Systems

- DeLite (Vor der Brück et al. 2008a,b) [Ankita Oswall]

22. Wednesday, January 25: Evaluation (Huenerfauth et al. 2009; van Oosten et al. 2010; Van Oosten et al. 2011) [Nika Strem]
23. Friday, January 27: Child Language Development
    - Revised D-Level (Lu 2009; Voss 2005) [Chrysanthi Melanou]
    - IPSyn (Sagae et al. 2005; Lubetich & Sagae 2014) [Neele Witte]
24. Wednesday, February 1: Reader Modeling
    - (Liu et al. 2004; Pendar & Chapelle 2008) [Peter Schoener]
    - (Nakatani et al. 2009, 2010) [Nikolas Zeitler]
25. Friday, February 3: Reader Modeling (cont.)
    - (Walmsley 2015) [Savvas Chatzipanagiiotidis]
26. Wednesday, February 8:
27. Friday, February 10:

Instructor: Detmar Meurers

- *Office:* Room 1.28, Blochbau (Wilhelmstr. 19)
- *Email:* dm@sfs.uni-tuebingen.de
- *Office hours:* Wednesdays 14:00–16:00 (arrange a slot by email beforehand)

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

- Wednesdays and Fridays, 8:30-10:00

Credit Points: 6 CP or 9 CP (with term paper)

- active participation in class: 4SWS * 15 = 60h (2 CP)
- reading and positing of questions: 60h (2 CP)
- preparing and holding class presentation: 60h (2 CP)
- optional: term paper 90h (3 CP)

Syllabus (this file):

- html-Version (http://purl.org/dm/16/ws/hs)
- pdf-Version (http://purl.org/dm/16/ws/hs/syllabus.pdf)

Moodle page:

- `https://moodle02.zdv.uni-tuebingen.de/course/view.php?id=1636`

Nature of course and our expectations: This is a research-oriented Hauptseminar, in which we jointly explore perspectives and approaches on complexity in linguistics, psycholinguistics, and computational linguistics. You are expected to

1. regularly and actively participate in class, read the papers assigned by any of the presenters and post a meaningful question on Moodle to the "Reading Discussion Forum" on each reading *at the latest on the day before it is discussed* in class.
2. explore and present a topic:
   - select one of the sub-topics by the end of October
   - thoroughly research the topic, taking our literature pointers *as a starting point*
   - prepare the presentation with slides, send them to me by email and discuss them with me in a half hour slot during my office hour *at least 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 and discuss the topic in class
3. if you pursue the 9 CP option, work out a project term paper
   - *by January 27, 2017* select a topic and submit a one-page abstract
     - For computational linguistics students, the topic of the paper will typically be the exploration or implementation of an approach analyzing complexity.
   - *by March 30, 2017, i.e. before the beginning of the next semester* email the term paper in pdf format to the instructor.
     - Note for Computational Linguistics students: The term paper must be produced in LaTeX using the ACL conference format or the Computational Linguistics journal format; 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. All portable electronic devices such as cell phones and laptops should be switched off for the entire length of the flight, oops, class.

References

Alexopoulou, T., M. Michel, A. Murakami & D. Meurers (submitted). Analyzing learner language in task contexts: A study case of task-based performance in EFCAMDAT. *Language Learning* Special Issue on "Language learning research at the intersection of experimental, corpus-based and computational methods: Evidence and interpretation".

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.

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://purl.org/net/Brown.Eskenazi-05.pdf`.

Cheung, H. & S. Kemper (1992). Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics* 13(01), 53–76. URL `http://dx.doi.org/10.1017/S0142716400005427`.

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.

Covington, M. A., C. He, C. Brown, L. Naçi & J. Brown (2006). *How complex is that sentence? A proposed revision of the Rosenberg and Abbeduto D-Level Scale*. Computer Analysis of Speech for Psychological Research (CASPR) Research Report 2006-01, The University of Georgia, Artificial Intelligence Center, Athens, GA. URL `http://www.ai.uga.edu/caspr/2006-01-Covington.pdf`.

Crossley, S. A., D. F. Dufty, P. M. McCarthy & D. S. Mcnamara (2000). Toward a New Readability : A Mixed Model Approach. In D. 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.

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://aclweb.org/anthology/W11-1409`.

Demberg, V. & F. Keller (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109(2), 193 – 210.

Demberg, V. & A. Sayeed (2011). Linguistic cognitive load: implications for automotive UIs. In *Adjunct Proceedings of AutomotiveUI'11*.

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*. Costa Mesa, California: Impact Information.

François, T. & E. Miltsakaki (2012). Do NLP and machine learning improve traditional readability formulas? In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*. Association for Computational Linguistics, pp. 49–57.

Graesser, A. C., D. S. McNamara, M. M. Louweerse & Z. Cai (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments and Computers* 36, 193–202. URL `http://home.autotutor.org/graesser/publications/bsc505.pdf`.

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.

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 at ACL-08*. Columbus, Ohio.

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, pp. 80–88.

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.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychological review* 95(2), 163.

Kintsch, W. & T. A. van Dijk (1978). Toward a Model of Text Comprehension and Productions. *Pschological Review* 85(5), 363–394.

Kyle, K. & S. A. Crossley (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly* 49(4), 757–786.

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.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition* 106 (3), 1126–1177.

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. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Languages Journal* pp. 190–208.

Lubetich, S. & K. Sagae (2014). Data-driven Measurement of Child Language Development with Simple Syntactic Templates. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, pp. 2151–2160. URL http://www.aclweb.org/anthology/C14-1203.

Malvern, D. D., R. B. J., C. N. & D. P. (2004). *Lexical diversity and language development: Quantification and assessment*. Palgrave Macmillan.

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. URL https://serifos.sfs.uni-tuebingen.de/svn/resources/trunk/papers/McCarthy.Jarvis-10.pdf.

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.

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.

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.

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.

Read, J. & P. Nation (2004). Measurement of formulaic sequences. *Formulaic sequences: Acquisition, processing and use* pp. 23–35.

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.

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.

Skehan, P. (1989). *Individual Differences in Second Language Learning*. Edward Arnold.

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.

van Schijndel, M. & W. Schuler (2016). Addressing surprisal deficiencies in reading time models. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC) at COLING*. Osaka.

Vor der Brück, T., S. Hartrumpf & H. Helbig (2008a). A Readability Checker with Supervised Learning using Deep Syntactic and Semantic Indicators. *Informatica* 32(4), 429–435.

Vor der Brück, T., H. Helbig & J. Leveling (2008b). *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* .

Walmsley, M. (2015). Learner Modelling for Individualised Reading in a Second Language. Ph.D. thesis, The University of Waikato. URL `http://hdl.handle.net/10289/10559`.

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.