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. Relatively, 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 [IULIJA 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 (Heliman et al. 2008a) [SOWMYA V.B.]
20. Wednesday, December 21:
21. Monday, January 9: T13 current approaches: Feng et al. [SOWMYA V.B.]
23. Monday, January 16: T14 current approaches: DeLite [MICHAEL HAHN]
24. Wednesday, January 18:
25. Monday, January 23: T15 Simplification Intro and early work [KAIKI 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)
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)

∗ 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
  ∗ T11 (Schwarm & 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 Welsley’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, 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


Petersen, S. E. & M. Ostendorf (2009). A machine learning approach to reading level


