Computational Approaches to Text Simplification

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.

For several of these notions of complexity, researchers have raised the question how a given text or sentence could be simplified. This includes a range of application contexts, from shortening sentences to make parsing more efficient to making texts accessible to children, second language learners, or people with disabilities.

In this seminar, we will discuss the empirical and conceptual nature of these notions of complexity and explore computational approaches to text simplification building on these.

Instructors:

• Detmar Meurers
  – *Office*: Room 1.28, Blochbau (Wilhelmstr. 19)
  – *Email*: dm@sfs.uni-tuebingen.de
  – *Office hours*: Mondays 18:00 (best to email beforehand)

• Sowmya V. B.
  – *Office*: Room 1.29, Blochbau (Wilhelmstr. 19)
  – *Email*: sowmya@sfs.uni-tuebingen.de
  – *Office hours*: arrange by email

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

• Mondays and Wednesdays, 16ct-18

Credits:

• Credit Points: 10 (MA ISCL)
• Credit Points: 15 (including BA thesis for BA ISCL)

Syllabus (this file):

• [html-Version](http://purl.org/dm/13/ss/simplification)
• [pdf-Version](http://purl.org/dm/13/ss/simplification/syllabus.pdf)
Moodle page: https://moodle02.zdv.uni-tuebingen.de/course/view.php?id=470

Nature of course and our expectations: This is a Hauptseminar (and BA thesis seminar) which on the one hand intends to provide an overview of current perspectives and approaches and on the other hand, offers computational linguistics students the opportunity to define and implement a text simplification approach, which typically forms the basis of the term paper or BA thesis project.

Students enrolled in the course are expected to

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)

   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):
   
   • 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
   • 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
   • prepare the presentation with slides and send them to both instructors at least two days before the 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 July 22
     – 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 both instructors before the beginning of the next semester, i.e., by September 30, 2013.
     – Note for ISCL students: The term paper must be produced in LaTeX using the Computational Linguistics journal style (http://cljournal.org/style.html), 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:** In principle, most of this is obvious, but to be explicit and clear: 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. All portable electronic devices such as cell phones should be switched off for the entire length of the flight, oops, class. Laptops should not be open in class unless there is a concrete, assigned activity. 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:**

1. **Introduction**
   What is readability? What is simplification? Applications, Motivations, Issues
   - Reading assignments:
     - April 22: (Vajjala & Meurers 2012)
     - April 24/29: (Feng 2008)
   Corpus studies, comparison of easy and difficult texts, psycholinguistic perspective on sentence comprehension
   - Reading assignments: (Allen 2009b)
   - Other relevant reading: (Petersen & Ostendorf 2007; Allen 2009a; Blum & Levenston 1978; Fukazawa 1994; Bautista 2010)
   - Chap 10 in (Aitchison 2011)
   - Simple Wiki guidelines\(^1\)
   - (Bautista et al. 2011)

2. **Background: Early Work**
   - Text simplification for Machines (Chandrasekar et al. 1996; Chandrasekar & Srinivas 1996)
   (Devlin 1999; Canning 2002) are PhD theses on this topic, which are not freely available online.

3. **A real-world application: The Porsimples project** (Aluíso et al. 2008; Candido et al. 2009; Aluíso & Gasperin 2010a,b; Belder & Moens 2010; Gasperin et al. 2009a)

4. **Lexical Simplification:**

• Corpus creation (De Belder & Moens 2012)
• SemEval 2012 Lexical Simplification task systems (Specia et al. 2012, . . .)
• Lexical Simplification via WordNet (Thomas & Anderson 2012) and via synonym replacement (Keskisärkkä 2012, for Swedish)
• A relatively full-fledged system for Spanish Lexical Simplification (Bott et al. 2012a; Drndarevic & Saggion 2012b,a; Drndarevic et al. 2012)
• More Advanced Lexical Simplification (Yatskar et al. 2010; Biran et al. 2011)

5. Syntactic Simplification

• Introduction (Gasperin et al. 2010)
• Coh-Metrix-based research (Crossley et al. 2012)
• Syntactic simplification with rules (Seretan 2012; Junior et al. 2011; Siddharthan 2011; Aranzabe et al. 2012b)
• Syntactic simplification using statistical machine translation (Specia 2010; Zhu et al. 2010; Coster & Kauchak 2011a,b; Woodsend & Lapata 2011a,b; Bach et al. 2011)


7. Identifying targets for simplification (Gasperin et al. 2009b; Medero & Ostendorf 2011; Bott & Saggion 2011a; Štajner, Drndarevic & Saggion 2013)

8. Corpus Creation (and systems) for other languages

• Basque (Aranzabe et al. 2012a)
• Brazilian Portuguese (Caseli et al. 2009)
• Danish (Klerke 2012; Klerke & Søgaard 2012)
• Italian (Barlacchi & Tonelli 2013)
• Spanish (Bott & Saggion 2011b; Bott et al. 2012b)
• Swedish (Smith & Jönsson 2011; Keskisärkkä 2012)

9. Application to other fields:

• Question Generation (Heilman & Smith 2010)
• Relation Extraction (Miwa et al. 2010)
• Spoken Language Understanding (Tur et al. 2011)
• Information Extraction (Jonnalagadda & Gonzalez 2010; J.Evans 2011)
• Ontology Mapping (Lu & Parameswaran 2009)
• Subtitling (Daelemans et al. 2004; Bouayad-Agha et al. 2006)
• Biomedical (Jonnalagadda & Gonzalez 2009; Jonnalagadda et al. 2009)
• Medical Literature (Damay et al. 2006)
• Crisis Management Domain (Temnikova 2012)
• Summarization for Simplification and Simplification for Summarization (Lal & Röger 2002; Blake et al. 2007; Margarido et al. 2008; Smith & Jönsson 2011; Advaith Siddharthan & McKeown 2004)

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, April 17: Organization and Introduction [Detmar Meurers]
2. Monday, April 22: Overview [Sowmya V.B.]
3. Wednesday, April 24: cont.
5. **Wednesday, May 1: Labor Day (the real one)**
7. **Wednesday, May 8: LEAD grad school**
9. Wednesday, May 15: TBD [NN]
10. **Monday, May 20: Pentecost holiday**
12. Monday, May 27: TBD [NN]
13. Wednesday, May 29: TBD [NN]
14. Monday, June 3: TBD [NN]
15. Wednesday, June 6: TBD [NN]
16. Monday, June 3: TBD [NN]
17. Wednesday, June 6: TBD [NN]
18. Monday, June 10: Project session
19. Wednesday, June 12: Project session
20. Monday, June 17: TBD [NN]
21. Wednesday, June 19: TBD [NN]
22. Monday, June 24: TBD [NN]
23. Wednesday, June 26: TBD [NN]
24. Monday, July 1: TBD [NN]
25. Wednesday, July 3: TBD [NN]
26. Monday, July 8: TBD [NN]
27. Wednesday, July 10: TBD [NN]
28. Monday, July 15: TBD [NN]
29. Wednesday, July 17: TBD [NN]

Last update: April 28, 2013
References


Belder, J. D. & M.-F. Moens (2010). Text Simplification For Children. In SIGIR workshop


Drndarevic, B., S. Štajner & H. Saggion (2012). Reporting Simply: A Lexical Simplifi-


