ISCL Hauptseminar (Sommersemester 2008)

Detmar Meurers: Using Natural Language Processing to Foster Language Awareness in Second Language Learning

Abstract: Complementing the general focus on communication and culture in foreign language teaching, a growing body of research since the 90s has established that awareness of language categories, forms and rules is important for an adult learner to successfully acquire a foreign language. Computer-Assisted Language Learning (CALL) systems could be one way to address this issue, but traditional CALL systems lack the ability to analyze language and provide feedback or input enhancement on that basis.

In this seminar we explore how natural language processing can be used to identify and represent relevant linguistic properties to overcome this shortcoming. The discussion of background and current research papers will be combined with group projects extending a prototype system (WERTi) that supports on-the-fly generation of language awareness materials based on web pages selected by the learner.

Topics and Issues:

- Introduction (Detmar):
  - Awareness and noticing of linguistic categories and forms
  - WERTi research: generating activities from authentic, learner-selected texts
- Obtaining texts: How can learners obtain authentic texts a) which they are interested in and thus motivated to read or work with, and b) which are appropriate for the learner level? How can one automatically measure 'coherence', 'complexity', 'difficulty' of a text?
  - Predicting reading difficulty Shizuka (1998); Schwarm & Ostendorf (2005); Heilman et al. (2008a); Kotani et al. (2008)
  - Lexical aspects, related to General Service List (GSL) and the Academic Word List (AWL). Cf. 'Vocabulary Resources' at `http://jbauman.com/`.
  - Automatic measurement of syntactic complexity
    - cf. articles on predicting reading difficult above
    - using the revised developmental scale Lu (2008b, under review)
    - Lu (2008a)
  - Coh-Metrix: `http://cohmetrix.memphis.edu/cohmetrixpr/index.html`
  - Automatic evaluation of IPSYN: Sagae et al. (2005), `http://www.cs.cmu.edu/~sagae/`
  - Automatically finding good examples Kilgarriff et al. (to appear in 2008); Segler (2007)
- How can authentic text material be presented or turned into activities?

  - Which linguistic forms, categories, or structures are relevant for language awareness in second language learning?
  - Which type of presentation or interaction is effective (or not) in fostering language awareness? Petersen (2005)
    * related to vocabulary acquisition (Sankó 2006)
  - Examples/Experiments with language awareness activities Matula (2007)
  - Which of those linguistic phenomena or properties can be reliably identified automatically? How, how reliably, and what impacts do errors made by the NLP technology have?
  - Which of those linguistic phenomena or properties, once identified, can be turned into enhanced presentations or activities? How reliably (e.g., which naturally occurring passives can be transformed to active sentences to ask learner to turn them back into the pragmatically appropriate original form)?
  - In case of activities requiring learner input, what feedback can automatically be provided to the learners for the automatically generated activities?

- Related work:
  - VISL: Visual Interactive Syntax Learning Bick (2005a,b) (Kilian)
  - Automatic generation of multiple choice "cloze tests" (FIB), typically for language testing and vocabulary drill Coniam (1997); Irvine & Kyllonen (2002); Deane & Sheehan (2003); Liu et al. (2005a); Huang et al. (2005); Liu et al. (2005b); Sumita et al. (2005); Smith et al. (2008a,b)
  - Exercise Authoring Tools (Ramon)
    * The Task Generator for ESL Toole & Heift (2001, 2002)
    * MIRTO Antoniadis et al. (2004)
    * ALFALEX: Automatic or semi-automatic generation of contextualized vocabulary exercises using a tagger and a parser. Verlinde et al. (2004)
  - Text retrieval for language practice (Katya):
    * REAP: Retrieval of texts for vocabulary and reading practice Heilman et al. (2008b,a)
    * Read-X Miltsakaki & Troutt (2008)
    * Finding texts for L2 learners of Chinese, Russian, English Sharoff et al. (2008)
  - Reading support tools (Desi):
    * Glosser Nerbonne & Smit (1996); Nerbonne et al. (1997, 1998)
    * COMPASS Breidt & Feldweg (1997)

- Lab:
  - Introduction to Web programming using WERTi as example (Detmar)
  - Extending WERTi with other text sources, text filtering, linguistic targets, and activity types (Everyone)

– Possible extra credit project: reimplementing WERTi as servlet using Spring (or related) web framework and UIMA backend.

I'm in contact with a lecturer in charge of teaching ESL classes at OSU, who would be happy to try out our prototypes with actual ESL learners.

Schedule:

- Thu, April 18.: Vorbesprechung
- Tue, April 22–May 6.:
  - *Topic:* Introduction to topic, own research, context Research
  - *Presenter:* Detmar
  - *Handout:* `http://purl.org/dm/08/ss/handouts/detmar-2x2.pdf`
- Thu, May 8.
  - *Topic:* Input Enhancement in Second Language Acquisition Research
  - *Presenter:* Magdalena
  - *Reading:* Introduction and Section 1 of Petersen (2005)
  - *Handout:* `http://purl.org/dm/08/ss/handouts/magdalena-2x2.pdf`
- Fri, May 9. Python/ModPython Tutorial
  - Homework Sheet 1: `http://purl.org/dm/08/ss/ex1.pdf`
- *Tue/Thu, May 13/15. no class (Pfingstferien)*
- Tue, May 20.
  - *Topic:* Input Enhancement in Second Language Acquisition Research (cont.)
  - *Presenter:* Magdalena
  - *Reading:* Sections 2 to end of Petersen (2005)
- *Thu, May 22. no class (Fronleichnam)*
- Fri, May 23. Python/ModPython Tutorial
  - Homework Sheet 2: `http://purl.org/dm/08/ss/ex2.pdf`
- Tue, May 27.
  - *Topic:* Language Activities in the VISL project
  - *Presenter:* Kilian
  - *Reading:* Bick (2005a,b)
  - *Handout:* `http://purl.org/dm/08/ss/handouts/kilian.pdf`
- Thu, May 29.
  - *Topic:* Coh-Metrix
  - *Presenter:* Maria
  - *Handout:* `http://purl.org/dm/08/ss/handouts/maria.pdf`
  - *Reading:*

    - Coh-Metrix: Analysis of text on cohesion and language `http://home.autotutor.org/graesser/publications/bsc505.pdf`
    - Validating Coh-Metrix `http://csep.psyc.memphis.edu/mcnamara/pdf/fpo444-mcnamara.pdf`

- Fri, May 30. Python/ModPython Tutorial
  - Homework Sheet 3: `http://purl.org/dm/08/ss/ex3.pdf`
- Tue, June 3.
  - *Topic:* Predicting Reading Difficulty
  - *Presenter:* Emma
  - *Reading:* Kotani et al. (2008); Heilman et al. (2008a); and as background, ch. 4 of Segler (2007), especially p. 46

    Links to some of the cited work: Shizuka (1998); Schwarm & Ostendorf (2005)
- Thu, June 5.
  - *Topic:* Automatic measurement of syntactic complexity using the revised developmental scale
  - *Presenter:* Tatiana
  - *Reading:* Lu (2008b, under review)
- Fri, June 6. Python/ModPython Tutorial
  - Homework Sheet 4: `http://purl.org/dm/08/ss/ex4.pdf`
- *Tue, June 10–19. (Detmar in Columbus as "local" ACL co-organizer, so no class – that's why we always start class 30 minutes early)*
- Tue, June 24.
  - *Topic:* Automatic evaluation of IPSyn
  - *Presenter:* Evgenia
  - *Reading:* Sagae et al. (2005) and as background for IPSyn Scarborough (1990) and as background comparing different scales Kemper et al. (1995)
- Thu, June 26.
  - *Topic:* Text retrieval for language practice
  - *Presenter:* Katya
  - *Reading:* Heilman et al. (2008b); Miltsakaki & Troutt (2008)
- Tue, July 1.
  - *Topic:* On finding good examples
  - *Presenter:* Niels

  - *Reading:* Kilgarriff et al. (to appear in 2008), and as background Church & Hanks (1990) and Kilgarriff (2006). See also Laufer (2008)

- Thu, July 3.
  - *Topic:* Exercise Authoring or Generation in MIRTO, ALFALEX, ESL Task Generator
  - *Presenter:* Ramon
  - *Reading:* Verlinde et al. (2004); Antoniadis et al. (2004); Toole & Heift (2001)

- Tue, July 8.
  - *Topic:* Automatic generation of multiple choice "cloze tests" (FIB)
  - *Presenter:* Aleks
  - *Reading:* `http://en.wikipedia.org/wiki/cloze`, Smith et al. (2008b), Liu et al. (2005b,a), Deane & Sheehan (2003)

- Thu, July 10.
  - *Topic:* Reading Support tools: GLOSSER, COMPASS
  - *Presenter:* Desi
  - *Reading:* Nerbonne et al. (1998); Breidt & Feldweg (1997)

- Tue, July 15.

- Thu, July 17.

References

Antoniadis, G., S. Echinard, O. Kraif, T. Lebarbé, M. Loiseau & C. Ponton (2004). NLP-based scripting for CALL activities. In E. H. Lothar Lemnitzer, Detmar Meurers (ed.), *COLING 2004 eLearning for Computational Linguistics and Computational Linguistics for eLearning*. Geneva, Switzerland: COLING, pp. 18–25. URL `http://acl.ldc.upenn.edu/coling2004/W6/pdf/3.pdf`.

Bick, E. (2005a). Grammar for Fun: IT-based Grammar Learning with VISL. In P. Juel (ed.), *CALL for the Nordic Languages*, Copenhagen: Samfundslitteratur, Copenhagen Studies in Language, pp. 49–64. URL `http://beta.visl.sdu.dk/pdf/CALL2004.pdf`.

Bick, E. (2005b). Live use of Corpus data and Corpus annotation tools in CALL: Some new developments in VISL. In H. Holboe (ed.), *Nordic Language Technology, Arbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004 (Yearbook 2004)*, Copenhagen: Museum Tusculanum, pp. 171–186. URL `http://beta.visl.sdu.dk/pdf/corpus_and_CALL_form.pdf`.

Breidt, E. & H. Feldweg (1997). Accessing Foreign Languages with COMPASS. *Machine Translation* 12(1–2), 153–174. Special Issue on New Tools for Human Translators.

Burstein, J. & C. Leacock (eds.) (2005). *Proceedings of the Second Workshop on Building Educational Applications Using NLP*. Ann Arbor, Michigan: Association for Computational Linguistics. URL http://www.aclweb.org/anthology-new/W/W05/#0200.

Church, K. & P. Hanks (1990). Word Association Norms, Mutual Information, And Lexicography. *Computational Linguistics* pp. 22–29. URL http://www.aclweb.org/anthology-new/J/J90/J90-1003.pdf.

Coniam, D. (1997). A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *Computer Assisted Language Instruction Consortium* 14(2–4), 15–33.

Deane, P. & K. Sheehan (2003). Automatic item generation via frame semantics. Education Testing Service report. URL http://eric.ed.gov/ERICWebPortal/recordDetail?accno=ED480135.

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: Association for Computational Linguistics. URL http://www.aclweb.org/anthology-new/W/W08/W08-0909.pdf.

Heilman, M., L. Zhao, J. Pino & M. Eskenazi (2008b). Retrieval of Reading Materials for Vocabulary and Reading Practice. In *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications*. Columbus, Ohio: Association for Computational Linguistics. URL http://www.aclweb.org/anthology-new/W/W08/W08-0910.pdf.

Huang, S.-M., C.-L. Liu & Z.-M. Gao (2005). Computer-assisted item generation for listening cloze tests and dictation practice in English. In *Advances in Web-Based Learning – ICWL. Proceedings of the 4th Int. Conference on Web-based Learning*, Berlin, Heidelberg: Springer, no. 3583/2005 in Lecture Notes in Computer Science.

Irvine, S. & P. Kyllonen (eds.) (2002). *Item Generation for Test Development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kemper, S., K. Rice & Y.-J. Chen (1995). Complexity metrics and growth curves for measuring grammatical development from five to ten. *First Language* 15, 151–166.

Kilgarriff, A. (2006). Collocationaility (and how to measure it). In *Proceedings of EURALEX*. URL http://www.kilgarriff.co.uk/Publications/2006-K-ELX-colloc.doc.

Kilgarriff, A., M. Husák, K. McAdam, M. Rundell & P. Rychlý (to appear in 2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of EURALEX-08*.

Kotani, K., T. Yoshimi, T. Kutsumi, I. Sata & H. Isahara (2008). EFL Learner Reading Time Model for Evaluating Reading Proficiency. In *Proceedings of CICLING*. Haifa, Israel. URL http://www.springerlink.com/content/d320h3p212257jk7/fulltext.pdf.

Laufer, B. (2008). Corpus-based versus Lexicographer Examples in Comprehension and Production of New Words. In T. Fontenelle (ed.), *Practical Lexicography: A Reader*, Oxford: Oxford University Press, pp. 213–218. URL http://purl.org/dm/08/ss/lit/laufer-08.pdf. First published 1992.

Liu, C.-L., C.-H. Wang & Z.-M. Gao (2005a). Using lexical constraints for enhancing computer-generated multiple-choice cloze items. *Int. Journal of Computational Linguistics and Chinese Language Processing* 10. URL http://www.aclclp.org.tw/clclp/v10n3/v10n3a1.pdf.

Liu, C.-L., C.-H. Wang, Z.-M. Gao & S.-M. Huang (2005b). Applications of Lexical Information for Algorithmically Composing Multiple-Choice Cloze Items. In Burstein & Leacock (2005), pp. 1–8. URL http://www.aclweb.org/anthology/W/W05/W05-0201.

Lu, X. (2008a). Automatic Measurement of Syntactic Complexity in Second Language Acquisition. URL http://purl.org/net/icall/calico08/lu.pdf. Slides of talk presented at the CALICO-08 Pre-Conference Workshop on "Automatic Analysis of Learner Language: Bridging Foreign Language Teaching Needs and NLP Possibilities". San Francisco.

Lu, X. (2008b). Automatic measurement of syntactic complexity using the revised developmental scale. In *Proceedings of the 21st International Florida Artificial Intelligence Research Society Conference (FLAIRS-08)*. Coconut Grove, FL: AAAI Press.

Lu, X. (under review). Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics* .

Matula, S. (2007). Incorporating a Cognitive Linguistic Presentation of the Prepositions 'on', 'in' and 'at' in ESL Instruction: A quasi-experimental study. Ph.D. thesis, Georgetown University.

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*. Columbus, Ohio: Association for Computational Linguistics, pp. 89–97. URL http://www.aclweb.org/anthology/W/W08/W08-0911.

Nerbonne, J., D. Dokter & P. Smit (1998). Morphological Processing and Computer-Assisted Language Learning. *Computer Assisted Language Learning* 11(5), 543–559.

Nerbonne, J., L. Karttunen, E. Paskaleva, G. Proszeky & T. Roosmaa (1997). Reading More Into Foreign Languages. In *Fifth Conference on Applied Natural Language Processing*. Washington Marriot Hotel, Washington, DC, USA, pp. 00–00. URL http://www.aclweb.org/anthology/A97-1020.pdf.

Nerbonne, J. & P. Smit (1996). GLOSSER-RuG: in Support of Reading. In *COLING96: Proceedings of the 16th International Conference on Computational Linguistics*. Copenhagen: Center for Sprogteknologi, pp. 830–835. URL http://www.aclweb.org/anthology-new/C/C96/C96-2140.pdf.

Petersen, K. (2005). Input Enhancement in Second Language Acquisition Research: A Critical Review. Second Qualifying Paper, presented to The Department of Linguistics at Georgetown University.

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`.

Sankó, G. (2006). The Effects of Form- and Meaning-Focused Hypertextual Input Modification on L2 Vocabulary Acquisition and Retention. Ph.D. thesis, University of Debrecen, Hungary. URL `http://hdl.handle.net/2437/3763`.

Scarborough, H. S. (1990). Index of Productive Syntax. *Applied Psycholinguistics* 11(1), 1–22.

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: Association for Computational Linguistics, pp. 523–530. URL `http://www.aclweb.org/anthology/P/P05/P05-1065`.

Segler, T. M. (2007). Investigating the Selection of Example Sentences for Unknown Target Words in ICALL Reading Texts for L2 German. Ph.D. thesis, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh. URL `http://homepages.inf.ed.ac.uk/s9808690/thesis.pdf`.

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.

Shizuka, T. (1998). The Effects of Stimulus Presentation Mode, Question Type, and Reading Speed Incorporation on the Reliability/Validity of a Computer-based Sentence Reading Test. *JACET Bulletin* 29, 155–172. URL `http://www2.ipcku.kansai-u.ac.jp/~shizuka/pubs/papers/paper1998_5.html`.

Smith, S., S. Sommers & A. Kilgarriff (2008a). Automatic cloze generation: getting sentences and distractors from corpora. In *Proceedings of the 8th Teaching and Language Corpora Conference (TaLC 8)*. Lisbon.

Smith, S., S. Sommers & A. Kilgarriff (2008b). Learning Words Right with the Sketch Engine and WebBootCat: Automatic Cloze Generation from Corpora and the Web. In *Proceedings of the 25th International Conference of English Teaching and Learning 2008 International Conference on English Instruction and Assessment*. Lisbon. URL `http://www.ccu.edu.tw/fllcccu/2008EIA/English/C37.pdf`.

Sumita, E., F. Sugaya & S. Yamamoto (2005). Measuring Non-native Speakers' Proficiency of English by Using a Test with Automatically-Generated Fill-in-the-Blank Questions. In Burstein & Leacock (2005), pp. 61–68. URL `http://www.aclweb.org/anthology/W/W05/W05-0210`.

Toole, J. & T. Heift (2001). Generating Learning Content for an Intelligent Language Tutoring System. In *Proceedings of NLP-CALL Workshop at AI-ED*. San Antonio, Texas: 10th International Conference on Artificial Intelligence in Education, AI-ED, pp. 1–8.

Toole, J. & T. Heift (2002). Task-Generator: A Portable System for Generating Learning Tasks for Intelligent Language Tutoring Systems. In *Proceedings of ED-MEDIA 02, World Conference on Educational Multimedia, Hypermedia and Telecommunications, Charlottesville, VA*. AACE, pp. 1972–1978.

Verlinde, S., T. Selva & J. Binon (2004). ALFALEX, un environnement d'aide à l'apprentissage lexical du français langue étrangère. In *Technologies de l'Information et de la Connaissance dans l'Enseignement Supérieur et de l'Industrie (TICE 2004)*. Compiègne, UniversitÈ de Technologie de Compiègne. France. URL `http://edutice.archives-ouvertes.fr/docs/00/02/76/04/PDF/Selva_Verlinde.pdf`.

Python References:

- Tutorials:
  - Very first tutorial: `http://infohost.nmt.edu/tcc/help/pubs/lang/pytut/index.html`
  - Fast intro of the essentials: `http://www.poromenos.org/tutorials/python`
  - More in-depth tutorial: `https://docs.python.org/3/tutorial/index.html`
  - Collection of python links: `http://www.whoishostingthis.com/resources/python/`
- Quick references:
  - `http://rgruet.free.fr/PQR2.3.html`
  - `http://infohost.nmt.edu/tcc/help/pubs/python22/`
  - `http://www.python.org/doc/QuickRef.html`
- Manual:
  - Python Library Reference: `http://docs.python.org/lib/lib.html`

Organization: Instructor: Prof. Dr. Detmar Meurers

- *Email:* dm@ling.osu.edu
- *Web:* `http://purl.org/dm`
- *Office hours:* Tuesdays 16:30–17:30 in Room 1.27 (Bloch Bau, Wilhelmstr. 19)

When:

- Course: Tuesdays/Thursdays, 13:45st – 16 in Seminarraum 1.13 (Bloch Bau, Wilhelmstr. 19)
- Lab: Fridays 13:00st-14:00 in the Computerpool (Bloch Bau, second floor)

Course website: `http://purl.org/dm/08/ss`

Syllabus as PDF: `http://purl.org/dm/08/ss/syllabus.pdf`

Course email: dm-ss08 (in the department network, i.e. @sfs.uni-tuebingen.de)

This email will reach everyone involved in the course.

Course server: aticall.sfs.uni-tuebingen.de

The department user ids of the seminar participants collected in the Questionnaire will be added to a unix group which will have access to this server for the web-server based Übungen. For security reasons, you will get a separate home directory on this machine and access is restricted to within the department net.

(In case you wonder about the machine name: ATICALL = Authentic Text ICALL)

Anonymous feedback: If you have comments, complaints, or ideas you'd like to send me anonymously, you can use the web form at `http://purl.org/dm/feedback/` to do so. Please send me ordinary email for anything that you'd like to receive a reply to—there really is no way for me to find out who sent us something via the anonymous feedback form!

Nature of course and my expectations: This is a research-oriented seminar, i.e., each participant is expected to take an active role as a researcher. More concretely, each participant is expected to

1. regularly and actively participate in the class discussion (30% of grade)

   > Note: Following the rules of the Neuphilologische Fakultät, missing more than two course meetings unexcused, automatically results in failing the class.

2. explore and present a topic (30% of grade):
   - select a topic, schedule a meeting with me to discuss with me what you'll explore **before the end of April**
   - thoroughly research it taking my literature pointers as a starting point
   - prepare the presentation with slides and discuss the presentation with me **at least a week before your presentation**
   - after our meeting, email the class what they should read to prepare for your presentation **at least a week before your presentation**
   - present it in class
3. read the papers assigned by me or the presenters and post a question to the course list on each reading the day *before* it is discussed in class. (10% of grade)
4. work out a research idea or small software project related to the topic of this seminar and write a short paper about it, following the 8 page ACL paper style files and guidelines, to be handed in **no later than the end of September 08**. (30% of grade)

In terms of your time commitment, this means you should plan in about two hours of preparation for each hour of class. You'll need this time to properly do the general readings, research your specific topic, etc.

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.

Related to this, university rules require every document you submit for a grade to be accompanied by the following signed statement:

> Hiermit versichere ich, dass ich die vorgelegte Arbeit selbstständig und nur mit den angegebenen Quellen und Hilfsmitteln einschließlich des www und

anderer elektronischer Quellen angefertigt habe. Alle Stellen der Arbeit, die ich anderen Werken dem Wortlaut oder dem Sinne nach entnommen habe, sind kenntlich gemacht.