

Linguistics 485/700

Automatic Analysis of Learner Language for ICALL

Autumn 2008

Course goals In this seminar, we will focus on the automatic analysis of learner language, with the aim of developing insights and techniques for intelligent computer-aided language learning (ICALL) systems. To that end, we will investigate the question of how to detect and diagnose non-targetlike constructions in learner language, as well as how to provide an analysis for learner language more generally.

Investigating these questions can help determine the most appropriate uses of natural language processing (NLP) in the analysis of learner language. Specific questions to be addressed include the following: (first articulated for the [CALICO 2008 Workshop on the Automatic Analysis of Learner Language](#))

- Which properties of learner language are useful and relevant to obtain for Foreign Language Teaching and current Second Language Acquisition research?
- What annotation scheme or (error) taxonomy is appropriate for this and how do different annotation schemes compare?
- How reliably can errors and other properties of learner language be obtained automatically given the current state-of-the art in NLP?
- What is the impact of the specific properties of learner language on the (re)use of NLP technology? How does it impact performance and the potential use of such technology in foreign language teaching tools?

A good amount of practical analysis will be done in this course, and we will focus heavily on the usage of determiners and prepositions in learners of English, as much work has been done for them. The discussion, however, will be applicable to other languages and constructions.

Instructor: Markus Dickinson

Office: Memorial Hall (MM) 317

Phone: 856-2535

E-mail: md7@indiana.edu

Office hours: (at least for the first week)

T 11:00am

R 1:00pm

or by appointment

Meeting time: MW, 1:00-2:15pm

Classroom: Ballantine Hall (BH) 118

Course website: <http://jones.ling.indiana.edu/~mdickinson/08/700/>

Course notes will be posted to this website.

Credits: 3

Course prerequisites: Either some background in computational linguistics or applied linguistics (second language acquisition) is recommended, though not required.

Readings: There will be weekly readings for discussion, most of which are available online, or through the library online system. The others I will make available to you as needed. See the schedule and bibliography at the end for the full selection and see the course requirements below.

Course requirements:

- **Discussion:** Each student will have to prepare and lead the discussion at least once during the semester, on a topic of particular interest to them, and all students are expected to participate in the discussion every class.
 - About a week before you lead, you should schedule a meeting with me, in order to go over what you’re going to lead on and to determine which readings we’ll cover and on what specific days.
 - In addition to discussion leading, I would like you to have some activity, such as annotation of data or experimenting with a machine learning system, for the class to do.
- **Readings:** Every topic has associated readings with it, which may change depending on the interests of the students. Additionally, if you are presenting, you might find other readings which are necessary to give a more complete picture of the topic.
- **Assignments:** There will be some short (annotation & computational) assignments throughout the semester. In order to emphasize the cross-disciplinary nature of this seminar and to allow students to utilize their strengths, teams of students from different backgrounds will work on the computational assignments together.
- **Final project:** You will also have to work on a final project/paper at the end of the semester. Details of this will be given later in the semester.
 - Those taking the course at the 700 level will have more to do on the project than those taking the course at the 485 level.
- Breakdown:

Participation	25%
Discussion leading	25%
Assignments	10%
Final project	40%
- If you feel that I have given you incorrect or improper feedback/grading, please contact me outside of class. I will be happy to address your concerns.

Academic Misconduct: Academic misconduct is not allowed in this course. The Indiana University *Code of Student Rights, Responsibilities, and Conduct* (<http://dsa.indiana.edu/Code/>) defines academic misconduct as “any activity that tends to undermine the academic integrity of the institution . . . Academic misconduct may involve human, hard-copy, or

electronic resources . . . Academic misconduct includes, but is not limited to . . . cheating, fabrication, plagiarism, interference, violation of course rules, and facilitating academic misconduct” (II. G.1-6).

Students with Disabilities: Students who need an accommodation based on the impact of a disability should contact me to arrange an appointment as soon as possible to discuss the course format, to anticipate needs, and to explore potential accommodations.

I rely on Disability Services for Students for assistance in verifying the need for accommodations and developing accommodation strategies. Students who have not previously contacted Disability Services are encouraged to do so (812-855-7578; <http://www.indiana.edu/~iubdss/>).

Practical sessions We will have a variety of practical sessions (approx. 5), mostly scheduled outside of class, so that we can learn how to automatically analyze learner language for ourselves. The topics to be covered include:

- Corpus annotation & XML
- Scripting, data formatting
- Working with learner corpora
- NLP tools
- Machine learning

Topics

Topic	Readings
(I)CALL context	Nerbonne (2003); Heift and Schulze (2007, ch. 1-2) Heift and Nicholson (2001); Nagata (2002); Amaral and Meurers (2006, 2007); Dickinson and Herring (2008) Chapelle (2005); Dickinson et al. (to appear)
Corpora & Grammatical annotation	Leech (2004)
Learner corpora & error tagging	Granger (2003, 2004); Pendar and Chapelle (2008); Nesselhauf (2004); Myles (2005) Díaz-Negrillo and Fernández-Domínguez (2006), Tono (2000)
Error taxonomies, standardization, L1	Dodigovic (2005, p. 85-90, 177-188), Ellis (1994, ch. 2), Suri and McCoy (1993) Abuhakema et al. (2008); Jang et al. (2008); Oyama et al. (2008) Juozulynas (1994); Sanders (1991) Cowan et al. (2003) Pendar and Kosterina (2008); Heift (2008) Gass and Selinker (2001, ch.3) Milton and Chowdhury (1994)
General techniques for handling learner/ungrammatical input	Schneider and McCoy (1998); Menzel and Schroeder (1999); Vandeventer Faltin (2003, ch. 1 & 2); Vandeventer Faltin (2003, ch. 3); Reuer (2003); Foster and Vogel (2004); Wagner et al. (2007); Izumi et al. (2003, 2004)

Topic	Readings
Determiners, Prepositions, & Postpositions	Nagata et al. (2006); Han et al. (2006); Turner and Charniak (2007); Lee and Seneff (2006) Tetreault and Chodorow (2008); Chodorow et al. (2007); Lee and Knutsson (2008) Dickinson and Lee (submitted); de Ilarraza et al. (2008) De Felice and Pulman (2008, 2007); Eeg-Olofsson and Knutsson (2003), Gamon et al. (2008)
Content assessment, other error types	Bailey and Meurers (2008), Boyd and Meurers (2008) + TBA

Outline of Schedule:

Approximate Dates	Topic	Leader(s)
Sep. 3, 8, 10	ICALL context	Markus
Sep. 15	Corpora & Grammatical annotation	Markus
Sep. 17, 22, 24, 29	Learner corpora & Error tagging	
Oct. 1, 6, 8, 13, 15, 20, 22, 27	Error taxonomies	
Oct. 29, Nov. 3, 5, 10	General techniques	
Nov. 12, 17, 19, 24, Dec. 1, 3	Determiners, Prepositions, & Postpositions	
Dec. 8, 10	Other topics	
Mon., Dec. 15, 5-7pm	Project reports	All

Disclaimer This syllabus is subject to change. All important changes will be made in writing, with ample time for adjustment.

References

- Abuhakema, Ghazi, Anna Feldman and Eileen Fitzpatrick (2008). A New Arabic Interlanguage Database: Collection, Annotation, Analysis. Pre-CALICO Workshop on “Automatic Analysis of Learner Language: Bridging Foreign Language Teaching Needs and NLP Possibilities”. March 18-19, 2008. San Francisco, CA. <https://calico.org/p-378-Program.html>.
- Amaral, Luiz and Detmar Meurers (2006). Where does ICALL Fit into Foreign Language Teaching? CALICO Conference. May 19, 2006. University of Hawaii <http://www.ling.ohio-state.edu/icall/handouts/calico06-amaral-meurers.pdf>.
- Amaral, Luiz and Detmar Meurers (2007). Putting activity models in the driver’s seat: Towards a demand-driven NLP architecture for ICALL. EUROCALL. September 7, 2007. University of Ulster, Coleraine Campus. <http://www.ling.ohio-state.edu/icall/handouts/eurocall07-amaral-meurers.pdf>.
- Bailey, Stacey and Detmar Meurers (2008). Diagnosing Meaning Errors in Short Answers to Reading Comprehension Questions. In *The 3rd Workshop on Innovative Use of NLP for Building Educational Applications (ACL08-NLP-Education)*. Columbus, OH, pp. 107–115. <http://aclweb.org/anthology-new/W/W08/W08-0913.pdf>.

- Boyd, Adriane and Detmar Meurers (2008). On Diagnosing Word Order Errors. Pre-CALICO Workshop on “Automatic Analysis of Learner Language: Bridging Foreign Language Teaching Needs and NLP Possibilities”. March 18-19, 2008. San Francisco, CA. <https://calico.org/p-378-Program.html>.
- Chapelle, Carol (2005). Computer-assisted language learning. In Hinkel (2005), pp. 743–755.
- Chodorow, Martin, Joel Tetreault and Na-Rae Han (2007). Detection of Grammatical Errors Involving Prepositions. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*. Prague, Czech Republic, pp. 25–30. <http://www.aclweb.org/anthology/W/W07/W07-1604>.
- Cowan, Ron, Hyun Eun Choi and Doe Hyung Kim (2003). Four questions for error diagnosis and correction in CALL. *CALICO Journal* 20(3), 451–463. https://calico.org/html/article_288.pdf.
- De Felice, Rachele and Stephen Pulman (2007). Automatically Acquiring Models of Preposition Use. In *Proceedings of the 4th ACL-SIGSEM Workshop on Prepositions*. Prague, Czech Republic, pp. 45–50. <http://www.aclweb.org/anthology/W/W07/W07-1607>.
- De Felice, Rachele and Stephen Pulman (2008). A classifier-based approach to preposition and determiner error correction in L2 English. In *Proceedings of COLING-08*. Manchester.
- de Ilarraza, Arantza Díaz, Koldo Gojenola and Maite Oronoz (2008). Detecting erroneous uses of complex postpositions in an agglutinative language. In *Proceedings of COLING-08*. Manchester.
- Díaz-Negrillo, Ana and Jesús Fernández-Domínguez (2006). Error Tagging Systems for Learner Corpora. *RESLA* 19, 83–102. dialnet.unirioja.es/servlet/fichero_articulo?codigo=2198610&orden=72810.
- Dickinson, Markus, Soojeong Eom, Yunkyoung Kang, Chong Min Lee and Rebecca Sachs (to appear). A Balancing Act: How can intelligent computer-generated feedback be provided in learner-to-learner interactions. *Computer Assisted Language Learning* .
- Dickinson, Markus and Joshua Herring (2008). Developing Online ICALL Exercises for Russian. In *The 3rd Workshop on Innovative Use of NLP for Building Educational Applications (ACL08-NLP-Education)*. Columbus, OH, pp. 1–9. <http://jones.ling.indiana.edu/~mdickinson/papers/dickinson-herring08.html>.
- Dickinson, Markus and Chong Min Lee (submitted). Modifying Corpus Annotation to Support the Analysis of Learner Language. *CALICO Journal* .
- Dodigovic, Marina (2005). *Artificial Intelligence in Second Language Learning: Raising Error Awareness*. Clevedon, UK: Multilingual Matters Ltd.
- Eeg-Olofsson, Jens and Ola Knutsson (2003). Automatic Grammar Checking for Second Language Learners - the Use of Prepositions. In *Proceedings of Nodalida'03*. Reykjavik, Iceland. <http://www.nada.kth.se/~knutsson/eegolofsson.knutsson.pdf>.
- Ellis, Rod (1994). *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

- Foster, Jennifer and Carl Vogel (2004). Parsing Ill-Formed Text Using an Error Grammar. *Artificial Intelligence Review: Special AICS 2003 Issue* 21, 269–291. http://www.computing.dcu.ie/~jfoster/publications/foster_aireview2004.pdf.
- Gamon, Michael, Jianfeng Gao, Chris Brockett, Alexander Klementiev, William Dolan, Dmitriy Belenko and Lucy Vanderwende (2008). Using Contextual Speller Techniques and Language Modeling for ESL Error Correction. In *Proceedings of IJCNLP*. Hyderabad, India. <http://research.microsoft.com/nlp/publications/IJCNLP.pdf>.
- Gass, Susan M. and Larry Selinker (2001). *Second Language Acquisition: An Introductory Course*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Granger, Sylvaine (2004). Computer learner corpus research: current status and future prospects. In U. Connor and T. Upton (eds.), *Applied Corpus Linguistics: A Multidimensional Perspective*, Amsterdam & Atlanta: Rodopi, pp. 123–145.
- Granger, Sylviane (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal* 20(3), 465–480. https://calico.org/html/article_289.pdf.
- Han, Na-Rae, Martin Chodorow and Claudia Leacock (2006). Detecting Errors in English Article Usage by Non-Native Speakers. *Natural Language Engineering* 12(2), 115–129.
- Heift, Trude (2008). Goals and Challenges for the Standardization of Error Typologies in Parser-Based CALL. Pre-CALICO Workshop on “Automatic Analysis of Learner Language: Bridging Foreign Language Teaching Needs and NLP Possibilities”. March 18-19, 2008. San Francisco, CA. <https://calico.org/p-378-Program.html>.
- Heift, Trude and Devlan Nicholson (2001). Web delivery of adaptive and interactive language tutoring. *International Journal of Artificial Intelligence in Education* 12(4), 310–325. http://aied.inf.ed.ac.uk/members01/archive/vol_12/heift/paper.pdf.
- Heift, Trude and Mathias Schulze (2007). *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues*. Routledge.
- Hinkel, Eli (ed.) (2005). *Handbook of Research in Second Language Teaching and Learning*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Izumi, Emi, Kiyotaka Uchimoto and Hitoshi Isahara (2004). SST speech corpus of Japanese learners’ English and automatic detection of learners’ errors. *ICAME Journal* 28, 31–48. <http://icame.uib.no/ij28/Izumi.pdf>.
- Izumi, Emi, Kiyotaka Uchimoto, Toyomi Saiga, Thepchai Supnithi and Hitoshi Isahara (2003). Automatic Error Detection in the Japanese Learners’ English Spoken Data. In *Proceedings of ACL-03*. Sapporo, Japan, pp. 145–148. <http://aclweb.org/anthology-new/P/P03/P03-2026.pdf>.
- Jang, Seok Bae, Sun-Hee Lee and Sang kyu Seo (2008). Annotation of Korean Learner Corpora for Particle Error Detection. Pre-CALICO Workshop on “Automatic Analysis of Learner Language: Bridging Foreign Language Teaching Needs and NLP Possibilities”. March 18-19, 2008. San Francisco, CA. <https://calico.org/p-378-Program.html>.

- Juozulynas, V. (1994). Errors in the Composition of Second-Year German Students: An Empirical Study of Parser-Based ICALI. *CALICO Journal* 12(1), 5–17. <http://calico.org/journalarticles/Volume12/vol12-1/Juozulynas.pdf>.
- Lee, John and Ola Knutsson (2008). The Role of PP Attachment in Preposition Generation. In *Proceedings of CICLing 2008, 9th International Conference on Intelligent Text Processing and Computational Linguistics*. Haifa, Israel. http://groups.csail.mit.edu/sls/publications/2008/cicling_johnlee.pdf.
- Lee, John and Stephanie Seneff (2006). Automatic Grammar Correction for Second-Language Learners. In *INTER-SPEECH 2006*. Pittsburgh, PA, pp. 1978–1981. <http://groups.csail.mit.edu/sls/publications/2006/IS061299.pdf>.
- Leech, Geoffrey (2004). Adding Linguistic Annotation. In Martin Wynne (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, Oxford: Oxbow Books, pp. 17–29. <http://ahds.ac.uk/linguistic-corpora/>.
- Menzel, Wolfgang and Inge Schroeder (1999). Error diagnosis for language learning systems. *Re-CALL* pp. 20–30. <http://nats-www.informatik.uni-hamburg.de/~ingo/papers/recall99.pdf.gz>.
- Milton, John C. P. and Nandini Chowdhury (1994). Tagging the interlanguage of Chinese learners of English. In *Proceedings joint seminar on corpus linguistics and lexicology, Guangzhou and Hong Kong, 19-22 June, 1993*. Language Centre, HKUST, Hong Kong, pp. 127–143. <http://repository.ust.hk/dspace/handle/1783.1/1087>.
- Myles, Florence (2005). Interlanguage corpora and second language acquisition research. *Second Language Research* 21(4), 373–391.
- Nagata, Noriko (2002). BANZAI: An Application of Natural Language Processing to Web based Language Learning. *CALICO Journal* 19(3), 583–599. <http://www.usfca.edu/japanese/CALICO02.pdf>.
- Nagata, Ryo, Atsuo Kawai, Koichiro Morihiro and Naoki Isu (2006). A Feedback-Augmented Method for Detecting Errors in the Writing of Learners of English. In *Proceedings of ACL-COLING-06*. Sydney, Australia, pp. 241–248. <http://www.aclweb.org/anthology/P06-1031>.
- Nerbonne, John (2003). Natural language processing in computer-assisted language learning. In Ruslan Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford University Press. <http://www.let.rug.nl/~nerbonne/papers/nlp-hndbk-call.pdf>.
- Nesselhauf, Nadja (2004). Learner corpora: Learner corpora and their potential for language teaching. In John McH. Sinclair (ed.), *How to Use Corpora in Language Teaching*, John Benjamins, pp. 125–152.
- Oyama, Hiromi, Yuji Matsumoto, Masayuki Asahara and Kosuke Sakata (2008). Construction of an error information tagged corpus of Japanese language learners and automatic error detection. Pre-CALICO Workshop on “Automatic Analysis of Learner Language: Bridging Foreign Language Teaching Needs and NLP Possibilities”. March 18-19, 2008. San Francisco, CA. <https://calico.org/p-378-Program.html>.

- Pendar, Nick and Carol Chapelle (2008). Investigating the Promise of Learner Corpora: Methodological Issues. *CALICO Journal* 25(2), 189–206. https://calico.org/html/article_689.pdf.
- Pendar, Nick and Anna Kosterina (2008). The Challenges of Annotating Learner Language. Pre-CALICO Workshop on “Automatic Analysis of Learner Language: Bridging Foreign Language Teaching Needs and NLP Possibilities”. March 18-19, 2008. San Francisco, CA. <https://calico.org/p-378-Program.html>.
- Reuer, Veit (2003). Error recognition and feedback with Lexical Functional Grammar. *CALICO Journal* 20(3), 497–512. <http://www.cl-ki.uni-osnabrueck.de/~vreuer/publ/calico03.reuer.pdf>.
- Sanders, R. H. (1991). Error Analysis in Purely Syntactic Parsing of Free Input. The Example of German. *CALICO Journal* 9(1), 72–89. <http://calico.org/journalarticles/Volume9/vol9-1/Sanders.pdf>.
- Schneider, David A. and Kathleen F. McCoy (1998). Recognizing Syntactic Errors in the Writing of Second Language Learners. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING) and the 36th Annual meeting of the ACL (ACL)*. Montreal, pp. 1198–1204. <http://acl.ldc.upenn.edu/P/P98/P98-2196.pdf>.
- Suri, L. Z. and K. F. McCoy (1993). *A methodology for developing an error taxonomy for a computer assisted language learning tool for second language learners*. Technical Report 93–16, Department of Computer and Information Sciences, University of Delaware, Newark, DE. <http://www.eecis.udel.edu/research/icicle/pubs/SuriMcCo93a.pdf>.
- Tetreault, Joel and Martin Chodorow (2008). The Ups and Downs of Preposition Error Detection in ESL Writing. In *Proceedings of COLING-08*. Manchester. <http://www.ets.org/Media/Research/pdf/r3.pdf>.
- Tono, Yukio (2000). A corpus-based analysis of interlanguage development: analysing POS tag sequences of EFL learner corpora. In *PALC'99: Practical Applications in Language Corpora..* pp. 323–340. <http://leo.meikai.ac.jp/~tono/paper/palc99.pdf>.
- Turner, Jenine and Eugene Charniak (2007). Language Modeling for Determiner Selection. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*. Rochester, New York, pp. 177–180.
- Vandeventer Faltin, Anne (2003). Syntactic error diagnosis in the context of computer assisted language learning. Thèse de doctorat, Université de Genève, Genève. <http://doc.rero.ch/getfile.py?docid=215&name=VandeventerA-these&format=pdf&version=1>.
- Wagner, Joachim, Jennifer Foster and Josef van Genabith (2007). A Comparative Evaluation of Deep and Shallow Approaches to the Automatic Detection of Common Grammatical Errors. In *Proceedings of EMNLP-CoNLL 2007*. pp. 112–121. http://www.computing.dcu.ie/~jfoster/publications/foster_emnlp2007.pdf.