On the Automatic Analysis of Learner Language. Introduction to the Special Issue

W. Detmar Meurers

Natural language processing (NLP) has long been used to automatically analyze language produced by language learners, typically aimed at providing individualized feedback and learner modeling in Intelligent Computer-Assisted Language Learning systems (cf. Heift & Schulze 2007). While much interesting research has been reported, it is difficult to determine the state of the art for the automatic analysis of learner language. Which error types and other learner language properties can be detected and diagnosed automatically? How reliably can this be done, for which kind of learner language, resulting from which types of tasks? For sustainable progress on the automatic analysis of learner language it arguably is crucial to answer these questions, to discuss and compare the performance of different analysis methods on real-life learner data sets.

As an essential prerequisite for addressing these issues, it is necessary to determine which learner language properties are useful or important to analyze in order to provide feedback and model language acquisition – a question which highlights the need for an intensive interdisciplinary dialogue between the fields of Intelligent Computer-Assisted Language Learning (ICALL), Second Language Acquisition (SLA), and Foreign Language Teaching (FLT).

Relatedly, the questions arising for the automatic analysis of learner language in ICALL intersect in important ways with research on learner corpora (cf. Granger 1998). Learner corpora in principle can help validate generalizations about language acquisition and provide a broad empirical basis for the development of new hypotheses and theories in SLA. However, to find the relevant classes of examples, the linguistic terminology used to single out the learner language aspects of interest needs to be mapped to instances in the corpus. Effective querying of corpora thus often requires reference to annotated linguistic abstractions instead of extensionally characterizing strings (cf. Meurers & Müller 2007).

Most of the work on annotating learner corpora has focused on the annotation of learner errors, for which a number of annotation schemes have been developed (Díaz-Negrillo & Fernández-Domínguez 2006). Yet, parallel to the ICALL situation mentioned above, there seems to be no agreement as to which distinctions are

needed and which can reliably and consistently be identified in learner language, be it manually or automatically; e.g., we are not aware of any studies reporting inter-annotator agreement figures for error annotation.[1]

Regarding the question which distinctions are useful or important to identify, SLA research essentially observes correlations of linguistic properties exhibited in learner language, whether erroneous or not. Correspondingly, the annotation of learner corpora should include a range of linguistic properties, including but not limited to learner errors. The challenge of defining linguistic annotation schemes for learner language and automatically annotating large learner corpora with such information has received little attention so far (the notable exceptions are de Haan 2000; van Rooy & Schäfer 2002, 2003; de Mönnink 2000), but there are encouraging signs from research on first language acquisition corpora, where recent work discusses the automatic analysis of morphological (MacWhinney 2008) and syntactic properties (Sagae et al. 2007; Lu 2009) and the use of CHILDES tools for SLA research (Myles & Mitchell 2004).

In sum, feedback and learner modeling in ICALL systems and the annotation of learner corpora for SLA and FLT research are both dependent on consistently identifiable learner language properties, their systematization in annotation schemes, and the development of NLP tools for automating such analysis as part of ICALL systems or to make the annotation of large learner corpora feasible. The papers collected in this special issue explore these issues further, by discussing the analysis of relevant aspects of written and spoken learner language, by defining and evaluating novel computational approaches, and by presenting systems integrating the analysis of learner language.

The idea for this special issue arose during the CALICO-08 pre-conference workshop on "Automatic Analysis of Learner Language: Bridging Foreign Language Teaching Needs and NLP Possibilities", where 30 talks and posters were presented March 18 and 19, 2008 in San Francisco (cf. `http://purl.org/calico/aall08.html` for abstracts and slides). The workshop brought together researchers working on the analysis of learner language in the broad sense, including work on annotation schemes for learner corpora and NLP techniques used to detect learner errors and other learner language properties. To further the discussion, we decided to organize a continuation, the "Automatic Analysis of Learner Language (AALL'09): From a better understanding of annotation needs to the development and standardization of annotation schemes", which took place March 10 and 11, 2009 at the Arizona State University in Tempe (cf. `http://purl.org/calico/`

[1]The one exception we are aware of is the Montclair Electronic Language Database (MELD), which has been annotated with reconstructions of the target forms. As discussed in Fitzpatrick & Seegmiller (2004), inter-annotator agreement was investigated and was found to be problematically low.

`aall09.html` for abstracts and slides) and to compile this special issue with selected papers.

Starting off the issue, in *Judging Grammaticality: Experiments in Sentence Classification*, the authors Joachim Wagner, Jennifer Foster, and Josef van Genabith investigate the general question how syntactically ill-formed sentences can be detected automatically. They discuss the nature of the information needed to automate such a classification, how to combine these sources of information, and for which cases such automatic classification is particularly successful.

In *Using Statistical Techniques and Web Search to Correct ESL Errors*, Michael Gamon, Claudia Leacock, Chris Brockett, William B. Dolan, Jianfeng Gao, Dmitriy Belenko, and Alexandre Klementiev discuss the motivation, design, and evaluation of the ESL assistant. The web-based system identifies typical learner errors and provides examples from the web for the original learner string and the suggested correction. The approach focuses on common problems in the English written by Chinese and Japanese learners, such as errors involving article, preposition, and auxiliary choice, overregularized verb inflection, noun number, local word order, as well as gerund/infinitive and adjective/noun confusions.

Rachele De Felice and Stephen Pulman zoom in on one of the most common errors made by learners of English by exploring the *Automatic Detection of Preposition Errors in Learner Writing*. They present the DAPPER system designed to recognize the obligatory use of nine prepositions and provide a detailed analysis of its performance on the Cambridge Learner Corpus.

A related, common cause of errors for learners of Korean are particles. Sun-Hee Lee, Seok Bae Jang, and Sang-Kyu Seo discuss the challenges involved in the *Annotation of Korean Learner Corpora for Particle Error Detection*. After presenting a classification of Korean particles and error types, they discuss the creation and particle error annotation of a Korean learner corpus. Based on this corpus, they provide an analysis of particle error types and error patterns, including a comparison of heritage and non-heritage learners.

In the paper *Modifying Corpus Annotation to Support the Analysis of Learner Language*, Markus Dickinson and Chong Min Lee also discuss the analysis of Korean particles. Here, however, the phenomenon serves to illustrate a general investigation into the relation between the phenomena found and annotations needed for learner corpora as compared to traditional native language corpora and NLP tools.

Returning from the analysis of specific phenomena to the design of complete systems analyzing and providing feedback to language learners, Noriko Nagata presents *Robo-Sensei: NLP-Based Error Detection and Feedback Generation*, an intelligent tutoring system for learners of Japanese. She describes the NLP components used and how they are combined to analyze and provide feedback to learner input, including all of the Japanese language structures introduced in the first two

years of a typical curriculum.

On the background of the intelligent tutoring system TAGARELA, for learners of Portuguese, Luiz Amaral and Detmar Meurers discuss *Little Things With Big Effects: On the Identification and Interpretation of Tokens for Error Diagnosis in ICALL*. Based on an analysis of the logs of learner interactions with the system, they discuss where mismatches between the learner conceptualization of tokens and the linguistic analysis performed by the system can lead to inappropriate feedback – and how an annotation-based NLP architecture can help address such mismatches in a general way.

In *Mastering Overdetection and Underdetection in Learner-Answer Processing: Simple Techniques for Analysis and Diagnosis*, the authors Alexia Blanchard, Olivier Kraif, and Claude Ponton highlight the importance of using NLP analysis in an appropriate didactic context. They show how such didactic triangulation supports high quality NLP analysis of learner language in activities generated as part of the ExoGen system.

Diane M. Napolitano and Amanda Stent present *TechWriter: An Evolving System for Writing Assistance for Advanced Learners of English*, a prototype writing assistant tool for advanced learners of English. The contribution emphasizes the importance of personalization, adapting the tool to the specific writer's weaknesses, and the importance of encouraging the writers to learn from their mistakes in order to foster writer autonomy.

The article *Computing Accurate Grammatical Feedback in a Virtual Writing Conference for German-Speaking Elementary-School Children: An Approach Based on Natural-Language Generation* by Karin Harbusch, Gergana Itsova, Ulrich Koch, and Christine Kühner also focuses on fostering writing skills, but it targets elementary-school children learning how to write essays in their native German. In so-called virtual writing conferences, the *Satzfee* system generates exercises in which learners use a drag-and-drop interface to compose stories, supported by system feedback.

Turning from written language to the spoken language produced by language learners, in *Annotation and Analyses of Temporal Aspects of Spoken Fluency* Heather Hilton provides a careful introduction to the methodology used in encoding temporal fluency phenomena in a spoken learner corpus, before presenting an exemplary comparison of the temporal structure of the speech of two subgroups of learners. The PAROLE corpus includes learners of French, English and Italian at various proficiency levels as well as native speakers.

Su-Youn Yoon, Lisa Pierce, Amanda Huensch, Eric Juul, Samantha Perkins, Richard Sproat, and Mark Hasegawa-Johnson report on the *Construction of a Rated Speech Corpus of L2 Learners' Spontaneous Speech*. The spontaneous speech corpus covers six language backgrounds and five proficiency levels and it is rated in

terms of general fluency score and phone accuracy, including comments on pronunciation errors.

Acknowledgements

Beginning with the AALL-08 workshop which got us started, I would like to thank Anne Rimrott as the co-organizer for her reliable and friendly help with the workshop. When we mentioned the idea of a special issue to Robert Fischer as the executive director of CALICO and main editor of the journal, he immediately was very encouraging. Throughout the process he has been a dedicated guide, putting in many extra hours to produce this issue. It has been a real pleasure working with him and Esther Horn – thank you very much for your kind and professional support!

Last but not least, I am grateful to the many reviewers who helped with selecting the papers for this issue and often provided substantial comments that helped the authors improve the contents and the writing: Luiz Amaral, Lars Borin, Kathy Corl, Ana Díaz Negrillo, Markus Dickinson, Eric Fosler, Jennifer Foster, Michael Gamon, Piklu Gupta, Trude Heift, Chris Hill, Emi Izumi, DJ Hovermale, Hye-Ri Joo, Ola Knutsson, Sun-Hee Lee, Sebastien L'Haire, Xiaofei Lu, Julie McGory, Lisa Michaud, Noriko Nagata, Martí Quixal, Sang Kyu Seo, Sylvie Thouësny, Yukio Tono, Martin Volk, Pauline Welby, Linda Wright. Thank you for your time and effort!

References

de Haan, P. (2000). Tagging non-native english with the tosca-icle tagger. In Mair & Hundt (2000), (pp. 69–79).

de Mönnink, I. (2000). Parsing a learner corpus. In Mair & Hundt (2000), (pp. 81–90).

Díaz-Negrillo, A., & Fernández-Domínguez, J. (2006). Error tagging systems for learner corpora. *Revista Española de Lingüística Aplicada (RESLA)*, *19*, 83–102.
URL http://dialnet.unirioja.es/servlet/fichero_articulo?codigo=2198610&orden=72810

Fitzpatrick, E., & Seegmiller, M. S. (2004). The montclair electronic language database project. In U. Connor, & T. Upton (Eds.) *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam: Rodopi.

URL http://chss.montclair.edu/linguistics/MELD/rodopipaper.pdf

Granger, S. (Ed.) (1998). *Learner English on Computer*. London; New York: Longman.

Heift, T., & Schulze, M. (2007). *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues*. Routledge.

Lu, X. (2009). Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, *14*, 3–28(26).
URL http://www.ingentaconnect.com/content/jbp/ijcl/2009/00000014/00000001/art00002

MacWhinney, B. (2008). Enriching childes for morphosyntactic analysis. In H. Behrens (Ed.) *Corpora in Language Acquisition Research. History, methods, perspectives*, vol. 6 of *Trends in Language Acquisition Research*, (pp. 165–198). Amsterdam and Philadelphia: John Benjamins.
URL http://childes.psy.cmu.edu/grasp/morphosyntax.doc

Mair, C., & Hundt, M. (Eds.) (2000). *Corpus Linguistics and Linguistic Theory*. Amsterdam: Rodopi.

Meurers, D., & Müller, S. (2007). Corpora and syntax (article 44). In A. Lüdeling, & M. Kytö (Eds.) *Corpus Linguistics. An International Handbook*, Handbooks of Linguistics and Communication Science. Berlin: Mouton de Gruyter.
URL http://purl.org/dm/papers/meurers-mueller-07.html

Myles, F., & Mitchell, R. (2004). Using information technology to support empirical sla research. *Journal of Applied Linguistics*, (pp. 169–196).

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2007). High-accuracy annotation and parsing of childes transcripts. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, (pp. 25–32). Prague, Czech Republic: Association for Computational Linguistics.
URL http://aclweb.org/anthology-new/W07-0604

van Rooy, B., & Schäfer, L. (2002). The effect of learner errors on pos tag errors during automatic pos tagging. *Southern African Linguistics and Applied Language Studies*, *20*, 325–335.

van Rooy, B., & Schäfer, L. (2003). An evaluation of three pos taggers for the tagging of the tswana learner english corpus. In D. Archer, P. Rayson, A. Wilson,

& T. McEnery (Eds.) *Proceedings of the Corpus Linguistics 2003 conference Lancaster University (UK), 28 – 31 March 2003*, vol. 16 of *University Centre For Computer Corpus Research On Language Technical Papers*, (pp. 835–844).