Haejoong Lee
haejoong@ldc.upenn.edu

## Education

University of Pennsylvania, Philadelphia
MSE, Computer and Information Science, May 2002

Sogang University, Seoul, Korea
MS, Computer Science, 1999
Construction of Korean Lexical Knowledge Base using Korean Machine Dictionary

Sogang University, Seoul, Korea
BS, Mathematics and Computer Science, 1997

## Skills

Python, Ruby, Perl, JavaScript, PHP, C, C++, Java, R
Ruby on Rails, Node.js, Java servlet
HTML5, CSS, HTTP, AJAX
XML, Unicode
MySQL, MongoDB, ElasticSearch
Android programming

## Software projects

Web transcription tool for Aikuma: A web based transcription tool with waveform display. Demo available at http://lp20.org/transcriber/test/web/transcriber.html. 2013~.

Toney: GUI tool for classifying spoken forms into phonetic categories. Written in C++ and Qt. http://lp20.org/toney/. 2013.

VScout: Firefox add-on for web video scouting. XUL, JavaScript with a backend built on Ruby on Rails. 2013.

Aikuma: Android app for recording spoken language. 2013~.

MCol: Live chat and SMS message collection system. 2012.

WebCol: Large-scale web screen-scraping framework. Python and MySQL. 2005.

XTrans: Multi-platform, multi-language linguistic transcription tool with audio waveform display and playback. C++, Qt, Python and PyQt.
https://www.ldc.upenn.edu/language-resources/tools/xtrans. 2005.

Open Language Archive Community: A virtual library of language resources. Main developer.
http://www.language-archives.org/. 2002~.

AGLIB: Annotation Graph Programming API written in C++. Participated as a file I/O developer. Wrote Python, Tcl, Java and Perl bindings. Developed file I/O plug-in architecture.
http://agtk.sourceforge.net/doc/aglib/2.0/. 2001.

**Work experience**

Aikuma: Development of speech recording app and supporting software infrastructure. 2013~.

DEFT Project: Data collection and annotation support. 2012~.

BOLT Project: Large scale web text collection and SMS and chat collection. 2012~.

TACKBP Project: Data collection and annotation support.

Usarufa Prosody: Analysis of Usarufa tone system using digital signal processing techniques and statistics. 2011.

RATS Project: Responsible for managing a programmer team who manages human annotation infrastructure and annotation data processing. 2011.

HAVIC Project: Responsible for annotation infrastructure. Developed video scouting tool called AScout and its backend. Developed web-based annotation tools. 2009~.

GALE Project: Developed a web harvesting framework and managed GALE web collection task. Developed a web data scouting tool called GScout that inspired other similar tool developments in LDC. Developed the LDC's next generation transcription tool called XTrans. Provide technical support for machine translation post editing task. Provide technical support for various aspects of the project. 2005-2001.

OLAC Project: Responsible for developing and managing the OLAC infrastructure built on web, XML and database technologies. 2002~.

ACE Project: Co-developed corpus exploration tool. Developed the early version of format converter between AG and ATF formats. Developed online annotation status reports. Involved

in data selection by processing raw HTML data. Provide technical support for various aspects of the project. 2003-2009.

QLDB Project: Implemented and tested the LPath+ query language, a language for querying linguistic database. Developed a linguistic database search tool called LPath QBA (Query By Annotation) based on the LPath+ technology. 2004-2007.

EARS Project: Designed data model/format and annotation tools for the Project. Took responsibilities in processing data and checking data integrity in entire data creation process. 2002-2005.

TalkBank Project: Developed Annotation Graph Toolkit. Developing GUI components of the new annotation toolkit. 2000-2004.

**Publications**

Steven Bird and Haejoong Lee. Computational support for early elicitation and classification of tone. Language Documentation & Conservation 8: 453—461, 2014.

Ann Bies, Zhiyi Song, Mohamed Maamouri, Stephen Grimes, Haejoong Lee, Jonathan Wright, Stephanie Strassel, Nizar Habash, Ramy Eskander and Owen Rambow. Transliteration of Arabizi into Arabi Orthography: Developing a Parallel Annotated Arabizi-Arabic Script SMS/Chat Corpus. In Proceedings of EMNLP 2014: Conference on Empirical Methods on Natural Language Processing, Doha, October 25-29, 2014.

Steven Bird, Florian R. Hanke, Oliver Adams and Haejoong Lee. Aikuma: A Mobile App for Collaborative Language Documentation. In Proceedings of ACL 2014: 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, June 22-27, 2014.

Zhiyi Song, Stephanie Strassel, Haejoong Lee, Kevin Walker, Jonathan Wright, Jennifer Garland, Dana Fore, Brian Gainor, Preston Cabe, Thomas Thomas, Brendan Callahan and Ann Sawyer. Collecting Natural SMS and Chat Conversations in Multiple Languages: The BOLT Phase 2 Corpus. In Proceedings of LREC 2014: 9th Edition of the Language Resources and Evaluation Conference, Reykjavik, May 26-31, 2014.

Jennifer Garland, Stephanie Strassel, Safa Ismael, Zhiyi Song and Haejoong Lee. Linguistic Resources for Genre-Independent Language Technologies: User-Generated Content in BOLT. In Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation, Istanbul, May 21-27, 2012.

Stephanie Strassel, Amanda Morris, Jonathan Fiscus, Christopher Caruso, Haejoong Lee, Paul Over, James Fiumara, Barbara Shaw, Brian Antonishek and Martial Michel. Creating HAVIC: Heterogeneous Audio Visual Internet Collection. In Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation, Istanbul, May 21-27, 2012.

Meghan Lammie Glenn, Stephanie Strassel, Haejoong Lee,Kazuaki Maeda, Ramez Zakhary and Xuansong Li. Transcription Methods for Consistency, Volume and Efficiency. In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May, 2010, Valletta, Malta, 2010.

Kazuaki Maeda, Haejoong Lee, Stephen Grimes, Jonathan Wright, Robert Parker, David Lee and Andrea Mazzucchi. Technical Infrastructure at Linguistic Data Consortium: Software and Hardware Resources for Linguistic Data Creation. In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta, 2010.

Meghan Lammie Glenn, Stephanie Strassel and Haejoong Lee. XTrans: a speech annotation and transcription tool. In Proceedings of INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009.

Kazuaki Maeda, Haejoong Lee, Shawn Medero, Julie Medero, Robert Parker and Stephanie Strassel. Annotation Tool Development for Large-Scale Corpus Creation Projects at the Linguistic Data Consortium. In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco, 2008.

Meghan Lammie Glenn, Stephanie Strassel, Lauren Friedman, Haejoong Lee and Shawn Medero, Management of Large Annotation Projects Involving Multiple Human Judges: a Case Study of GALE Machine Translation Post-editing. In Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco, 2008.

Lauren Friedman, Stephanie Strassel and Haejoong Lee. A Quality Control Framework for Gold Standard Reference Translations: The Process and Toolkit Developed for GALE Translating & The Computer 30 (hosted by EAMT): London, November 19-20, 2008.

Steven Bird and Haejoong Lee. Graphical Query for Linguistic Treebanks. In Proceedings of PACLING 2007 - 10th Conference of the Pacific Association for Computational Linguistics, pages pp. 22-30, Melbourne, 2007.

Kazuaki Maeda, Haejoong Lee, Julie Medero and Stephanie Strassel. A New Phase in Annotation Tool Development at the Linguistic Data Consortium: The Evolution of the

Annotation Graph Toolkit. In Proceedings of LREC 2006: 5th International Conference on Language Resources and Evaluation, Genoa, 2006.

Ann Bies, Stephanie Strassel, Haejoong Lee, Kazuaki Maeda, Seth Kulick, Yang Liu, Mary Harper and Matthew Lease. Linguistic Resources for Speech Parsing. In Proceedings of LREC 2006: 5th International Conference on Language Resources and Evaluation, Genoa, 2006.

Steven Bird, Yi Chen, Susan Davidson, Haejoong Lee and Yifeng Zheng. Designing and Evaluating an XPath Dialect for Linguistic Queries. In Proceedings 22nd International Conference on Data Engineering (ICDE), Atlanta, USA, 2006.

Steven Bird, Yi Chen, Susan Davidson, Haejoong Lee and Yifeng Zheng. Extending XPath to Support Linguistic Queries. In Proceedings of Programming Language Technologies for XML (PLANX), pp.35-46, Long Beach, California, 2005.

Baden Hughes, Steven Bird, Haejoong Lee and Ewan Klein. Experiments with data-intensive NLP on a computational grid. In Proceedings of the 2004 Hong Kong International Workshop on Language Technology, 2004.

Steven Bird, Kazuaki Maeda, Xiaoyi Ma, Haejoong Lee, Beth Randall and Salim Zayat. TableTrans, MultiTrans, InterTrans and TreeTrans: Diverse Tools Built on the Annotation Graph Toolkit. In Proceedings of the 3rd International Conference on Language Resources and Evaluation, Paris: European Language Resources Association, 2002.

Xiaoyi Ma, Haejoong Lee, Steven Bird and Kazuaki Maeda. Models and Tools for Collaborative Annotation. In Proceedings of the 3rd International Conference on Language Resources and Evaluation, Paris: European Language Resources Association, 2002.

Kazuaki Maeda, Steven Bird, Xiaoyi Ma and Haejoong Lee. Creating Annotation Tools with the Annotation Graph Toolkit. In Proceedings of the 3rd International Conference on Language Resources and Evaluation, Paris: European Language Resources Association, 2002.

Steven Bird, Kazuaki Maeda, Xiaoyi Ma and Haejoong Lee. Annotation Tools Based on the Annotation Graph API. In Proceedings of ACL/EACL Workshop on Sharing Tools and Resources, pp.31-44, Toulouse, 2001.

Kazuaki Maeda, Steven Bird, Xiaoyi Ma, Haejoong Lee. The Annotation Graph Toolkit: Software Components for Building Linguistic Annotation Tools. In Proceedings of HLT 2001: The Human Language Technologies Conference, San Diego, March, 2001.

Haejoong Lee, Jeongmi Cho and Jungyun Seo. Korean Lexical Knowledge Base Construction System. In Proceedings of 11th Conference on Korean and Korean Language Processing, pp.387-403, Seoul, Korea, 1999.

Harksu Kim, Haejoong Lee and Jungyun Seo. Analysis of Dialogue Discourse Structure Using Neural Network. In Proceedings of 15th Conference on Korean Speech Communication and Signal Processing (KSCSP'98), Volume 15, No.1, pp.419-424, Seoul, Korea, 1998.