David R. Mortensen

Language Technologies Institute · School of Computer Science · Carnegie Mellon University
April 24, 2021

Office: 5407 Gates Hillman Complex, 4902 Forbes Avenue, Pittsburgh, PA 15213, USA
Phone: +1 412 527 8827
Email: dmortens@cs.cmu.edu
Website: `http://cs.cmu.edu/~dmortens/`

Education

- PhD in Linguistics. 2006. University of California, Berkeley. Dissertation: *Logical and Substantive Scales in Phonology*, supervised by Sharon Inkelas and James A. Matisoff (co-chairs), Larry Hyman and Johanna Nichols (committee members).

- LSA Summer Institute. 2006. Massachusetts Institute of Technology.

- MA in Linguistics. 2003. University of California, Berkeley.

- BA in English Literature (Departmental Honors; minors in Linguistics and Psychology). 2000. Utah State University. Honors Thesis: *Sinitic Loanwords in Two Hmong Dialects of Southeast Asia*.

Professional Appointments

2018-present	Systems Scientist, Language Technologies Institute, Carnegie Mellon Univeristy
2015–2018	Senior Researcher, Language Technologies Institute, Carnegie Mellon University
2014–2014	Visiting Researcher, Language Technologies Institute, Carnegie Mellon University
2006–2014	Assistant Professor, Department of Linguistics, University of Pittsburgh

Publications

Journal Articles

Patrick Littell, Tian Tian, Ruochen Xu, Zaid Sheikh, David Mortensen, Lori Levin, Francis Tyers, Hiroaki Hayashi, Graham Horwood, Steve Sloto, Emily Tagtow, Alan Black, Yiming Yang, Teruko Mitamura, and Eduard Hovy. 2018b. The ARIEL-CMU situation frame detection pipeline for LoReHLT16: a model translation approach. *Machine Translation*, 32(1–2):105–126.

David R. Mortensen. 2013b. Tonally conditioned vowel raising in Shuijingping Mang. *Journal of East Asian Linguistics*, 22(3):189–216.

David R. Mortensen. 2012a. The emergence of obstruents after high vowels. *Diachronica*, 29(4):434–470.

David R. Mortensen and James A. Miller. 2013. A reconstruction of Proto-Tangkhulic rhymes. *Linguistics of the Tibeto-Burman Area*, 36(1):1–32.

Mary Lou Vercellotti and David R. Mortensen. 2012. A classification of compounding in American Sign Language: an evaluation of the Bisetto and Scalise framework. *Morphology*, 22(4):545–579.

David R. Mortensen and Jennifer Keogh. 2012. Sorbung, an undescribed language of Manipur: Its phonology and place in Tibeto-Burman. *Journal of the Southeast Asian Linguistics Society*, 4(1):62–114.

Refereed Conference Papers

Xinjian Li, David R Mortensen, Florian Metze, and Alan W. Black. 2021. Multilingual phonetic dataset for low resource speech recognition. In *ICASSP 2021*.

Jimin Sun, Hwijeen Ahn, Chan Young Park, Yulia Tsvetkov, and David R. Mortensen. 2021. Ranking transfer languages with pragmatically-motivated features for multilingual sentiment analysis. In *EACL 2021*.

Aditi Chaudhary, Antonios Anastasopoulos, Adithya Pratapa, David R. Mortensen, Zaid Sheikh, Yulia Tsvetkov, and Graham Neubig. 2020. Automatic extraction of rules governing morphological agreement. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Shahan Ali Memon, Aman Tyagi, David R. Mortensen, and Kathleen M Carley. 2020. Characterizing sociolinguistic variation in the competing vaccination communities. In *Proceedings of the International Conference SBP-BRiMS 2020*, Washington DC. Springer.

Maria Ryskina, Ella Rabinovich, Taylor Berg-Kirkpatrick, David R. Mortensen, and Yulia Tsvetkov. 2020. Where new words are born: Distributional semantic analysis of neologisms and their semantic neighborhoods. In *Proceedings of the Society for Computation in Linguistics*, volume 3.

David R. Mortensen, Xinjian Li, Patrick Littell, Alexis Michaud, Shruti Rijhwani, Antonios Anastasopoulos, Alan W. Black, Florian Metze, and Graham Neubig. 2020. AlloVera: A multilingual allophone database. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R. Mortensen, Graham Neubig, Alan W Black, and Florian Metze. 2020a. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020*.

Xinjian Li, Siddharth Dalmia, David R. Mortensen, Juncheng Li, Alan Black, and Florian Metze. 2020b. Towards zero-shot learning for automatic phonemic transcription. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Aditi Chaudhary, Chunting Zhou, Lori Levin, Graham Neubig, David R. Mortensen, and Jaime Carbonell. 2018. Adapting word embeddings to new languages with morphological and phonological subword representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3285–3295, Brussels, Belgium. Association for Computational Linguistics.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Patrick Littell, Tom McCoy, Na-Rae Han, Shruti Rijhwani, Zaid Sheikh, David Mortensen, Teruko Mitamura, and Lori Levin. 2018a. Parser combinators for Tigrinya and Oromo morphology. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Kartik Goyal, Chris Dyer, and Lori Levin. 2016b. Bridge-language capitalization inference in western Iranian: Sorani, Kurmanji, zazaki, and Tajik. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3318–3324, Portorož, Slovenia. European Language Resources Association (ELRA).

Yulia Tsvetkov, Sunayana Sitaram, Manaal Faruqui, Guillaume Lample, Patrick Littell, David Mortensen, Alan W Black, Lori Levin, and Chris Dyer. 2016. Polyglot neural language models: A case study in cross-lingual phonetic representation learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1357–1366, San Diego, California. Association for Computational Linguistics.

Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. 2016. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472, Austin, Texas. Association for Computational Linguistics.

Patrick Littell, Kartik Goyal, David R. Mortensen, Alexa Little, Chris Dyer, and Lori Levin. 2016a. Named entity recognition for linguistic rapid response in low-resource languages: Sorani Kurdish and Tajik. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 998–1006, Osaka, Japan. The COLING 2016 Organizing Committee.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.

Other Academic Publications

David R. Mortensen. 2019. Hmong (Mong Leng). In Alice Vittrant and Justin Watkins, editors, *The Mainland Southeast Asia Linguistic Area*. De Gruyter Mouton, Berlin, Boston.

David R. Mortensen. 2017a. Hmong-Mien languages. In Mark Aronoff, editor, *Oxford Research Encyclopedia of Linguistics*. Oxford University Press, New York.

David R. Mortensen. 2017b. Hmong-Mien languages. In Mark Aronoff, editor, *Oxford Bibliographies in Linguistics*. Oxford University Press, New York.

David R. Mortensen. 2013a. Lexical prefixes and Tibeto-Burman laryngeal contrasts. In *Proceedings of the Thirty-Seventh Annual Meeting of the Berkeley Linguistics Society (BLS 37)*, page 292–303.

David R. Mortensen. 2004. The emergence of dorsal stops after high vowels in Huishu. In *Proceedings of the Thirtieth Annual Meeting of the Berkeley Linguistics Society (BLS 30)*, pages 292–303.

David R. Mortensen. 2003. Review of *Baheng Yu Yanjiu* [research on the Pa-Hng language] by Mao Zongwu and Li Yunbing. *Linguistics of the Tibeto-Burman Area*, 26(1):123–131.

David R. Mortensen. 2002. Review of *Les langue Hmong-Mjen: phonologie historique* by Barbara Niederer. *Linguistics of the Tibeto-Burman Area*, 25(2):243–245.

Data Resources

Mortensen, David R. 2014/2019. SCH Corpus. 15 million word unlabeled Hmong-language corpus derived from postings to the `soc.culture.hmong` Usenet group.
`http://www.davidmortensen.org/corpora/sch_corpus-2.zip`

Software

David R. Mortensen and Jong Hyuk Park. 2019. IndoMorph. Collection of Foma FST morphological analyzers for languages of the Indian subcontinent.

David R. Mortensen. 2018b. MStem. Python multilingual morphological stemming framework and stemmer collection.

David R. Mortensen. 2018a. Epitran. Precision orthography-to-IPA conversion for 65 languages.

David R. Mortensen. 2016. PanPhon. Articulatory feature extractor and library.

David R. Mortensen. 2012b. NetSPE. A web-based application for exploring rule-based analyses of phonological data.

David R. Mortensen. 2011b. Web Comparator. A web-based application for organizing and analyzing comparative lexical databases. Used in the production of "Emergence of Obstruents" and "Proto-Tangkhulic Rhymes".

David R. Mortensen. 2011a. HsSPE. Haskell library implementing SPE-style rule-based phonology.

Selected Talks

David R. Mortensen. 2019. "Hmong Elaborate Expressions; Constructional and Distributional-Semantic Perspectives." colloquium talk, Carnegie Mellon University, April 19, 2019.

David R. Mortensen. 2019. "Hmong Elaborate Expressions; Constructional and Distributional-Semantic Perspectives." invited talk, University of Minnesota, Minneapolis, March 29, 2019.

David R. Mortensen. 2010. "Does Hmong allow noun incorporation? A field study," colloquium talk, University of Pittsburgh, November 19.

David R. Mortensen. 2010. "Does Hmong allow Noun Incorporation?" invited talk, Wayne State University, Detroit, October 1.

David R. Mortensen. 2010. "Shuijingping Hmong and the problem of tone-vowel interactions," invited talk, the University at Buffalo, February 12.

David R. Mortensen. 2009. "The Problem of Tone-Vowel Interactions," invited talk, Chulalongkorn University, Bangkok, November 10.

David R. Mortensen. 2008. "Optimal-Constructional Morphology, or how, when our lexicon gives us lemons, we make lemonade," colloquium talk, University of Pittsburgh, November 25.

David R. Mortensen. 2006. "A Marked Structure is Born: The Emergence of Obstruents and the Motivations for Sound Change," invited talk, University of Pennsylvania, Philadelphia, February 16.

David R. Mortensen. 2005. "The Phonology of Coordinate Constructions," invited talk, Cornell University, Ithaca, March 3.

David R. Mortensen. 2005. "Abstract Scales in Phonology," invited talk, Stanford Phonology Workshop, Stanford February 24.

David R. Mortensen. 2005. "Abstract Scales in Phonology," invited talk, University of Chicago, Chicago, Febraury 3.

David R. Mortensen. 2004. "Scalar Competition and Variables in Hmong Anaphora," invited talk, Stanford Syntax Workshop, Stanford, April 27, 2004. David R. Mortensen. 2003. "The State of Hmong Linguistics," invited talk, Hmong Studies Symposium, Concordia University (St. Paul), October 25, 2003.

Teaching

North American Summer School in Logic, Language, and Information (NASSLLI 2018)

Low Resource Techniques in NLP	Summer 2018

Carnegie Mellon University

11-624/11-724: Human Language for AI	F 2018, S 2019, F 2019, F 2020
11-411/11-611: Natural Language Processing	S 2017, F 2017, S 2018, F 2018, S 2019, F 2019, S 2020, F 2020, S 2021
11-421/11-721: Grammars and Lexicons	F 2017, F 2018

University of Pittsburgh

Ling 1441/2441: Field Methods	S 2007, 2008, 2010, 2011, F 2012, S 2012, S 2014
Ling 1578: Phonetics and Phonemics	F 2008, 2010, 2011
Ling 1579: Phonology	S 2007, 2010, 2013
Ling 1773: Morphology	F 2006
Ling 1951: Languages of the World	F 2007, 2008, 2011, 2013
Ling 2578: Phonetics and Phonemics	F 2006
Ling 2579: Phonology	S 2008, 2010, 2012, 2013, 2014
Ling 2580: Topics in Phonological Theory	S 2008
Ling 2773: Morphology	F 2007, 2012, 2013
Ling 3773: Advanced Morphology	S 2011

Advising

PhD Students

Wang Ping. 2015. Phonological prominence and its interaction with tone in Chinese dialects. PhD dissertation, University of Pittsburgh. (Co-chair: Shelome Gooden)

PhD Dissertation Committees

Marcos, Nausica (in progress). Instruction and Learning of Derivational Morphology by English and German-speaking Learners of Spanish as a Foreign Language. PhD dissertation. (Chair: Alan Juffs).

Friedline, Benjamin. 2011. Morphological Challenges in the Second Language Acquisition of Derivational Morphology. PhD dissertation. (Chair: Alan Juffs).

Rodrguez, Guillermo. 2008. Second language sentence processing: Is it fundamentally different? PhD dissertation. (Chair: Alan Juffs)

Eberhardt, Maeve. 2009. Identities and Local Speech in Pittsburgh: A Study of Regional African American English. (Chair: Scott Kiesling).

MA Students

Jong Hyuk Park. 2020. Learning Internal Structure of Words with Factored Generative Models. MA Thesis. (Chair).

Cui Jie. 2010. Variation in Nanchang Gan. MA Thesis. (Chair).

Draucker, Fawn. 2010. Revisiting the Maori English Vowel Space: Exploring variation in /I/ and /u/ vowel production in Auckland, New Zealand. MA thesis. (Chair).

Littell, Patrick. 2008. "Body part incorporation." MA long paper. (Co-chair with Pascual Masullo).

MA Thesis Committees

Fan, Ya-wen. 2013. Gender Influences on the Use of Mandarin Varieties in Taiwan. MA Thesis. (Chair: Scott Kiesling).

Friedline, Benjamin. 2008. A Linguistic Profile of Power and Identity in Massively Multiplayer Online Role-Playing Games. (Chair: Scott Kiesling)

Collister, Lauren B.. 2008. Virtual discourse structure: an analysis of conversation in World of Warcraft. MA thesis. (Chair: Scott Kiesling)

McCarthy, Natalie. 2007. The Northern Cities Shift and local identity in a suburban Cleveland group. MA thesis. (Chair: Scott Kiesling).

Comprehensive/Qualifying Paper Committees

Tse, Holman. 2013. "The emergence of retroflection in Somali Bantu Kizigua: Internal motivation or contact-induced change?" (Chair)

Wang, Ping. 2012. "The duration of medial glides in first and second language speakers of Mandarin."

Olsen, Michael. 2012. " The Effect of Animacy and Pronominal Object Clitic Distinction in L2 Spanish." (Chair: Alan Juffs).

Marcos, Nausica. "An Analysis of Spanish Verb-Noun Compounds as Input for Derivation." (Chair: Alan Juffs).

Wang, Ping. 2012. "An OT Account of Major Patterns of Chinese Coda Evolution." (Chair).

Vercellotti, Mary Lou. 2011. "A classification of compounding in American Sign Language and an evaluation of the Bisetto and Scalise Framework." Comprehensive paper. (Co-chair with Claude Mauk).

Luk, Pei-Su. 2011. "Investigating the transitive and intransitive constructions in English, Chinese, and Japanese from a functional perspective." (Chair: Yasuhiro Shirai).

Collister, Lauren B.. 2010. "Meaning Variation of the Iconic Deictics ˆ and <– in an Online Community." (Co-chair with Scott Kiesling).

Lifrieri, Vernica. 2009. "A puzzle for effort in phonetically-grounded OT: codas in North-Central Peninsular Spanish." Comprehensive paper. (Co-chair with Shelome Gooden).

Chan, Derek Ho-Leung. 2009. "The tonal morphology of Cantonese attenuative reduplication." Comprehensive paper. (Co-chair with Shelome Gooden).

Martelle, Wendy. 2007. "The evolution of aspect: a case study in Russian." Comprehensive paper. (Chair: Yasuhiro Shirai)

Pharmakis, Patricia. 2007. "A formal account of Modern Greek inflection within the framework of Distributed Morphology." Comprehensive paper. (Chair: Alan Juffs)

Brubaker, Brian. 2007. "Implications of labial-tone correspondence in L2 Mandarin Chinese." Comprehensive paper. (Co-chair with Shelome Gooden).

Fellowships and Awards

Fellowships

- 2006. LSA Summer Institute Fellowship.
- 2004. Graduate Division Summer Grant, University of California, Berkeley.
- 2001. Institute for East Asian Studies Fellowship, University of California, Berkeley.

Research Grants and Contracts

- (pending). Automatic Speech Recognition and Transfer Learning: Large Corpus Production of Five Eastern Nahuatl Languages (east2720) with end-to-end technology. NSF DEL. $240,000.
- 2020. KAIROS. DARPA contract. Key Personnel. $11,350,578.
- 2010. University of Pittsburgh Arts and Sciences Faculty Grants Committee Summer Research Grant, Type I.

Professional Service

Editorial Boards

- Journal of the Southeast Asian Linguistics Society

Reviewer

- **Conferences:** ACL (Association for Computational Linguistics), COLING, EMNLP (Empirical Methods in Natural Language Processing), LREC (Language Resource and Evaluation Conference), NAACL (North American Chapter of the Association of Computational Linguistics)
- **Journals:** Computational Linguistics, Himalayan Linguistics, Diachronica, Language and Intercultural Communication, Lingua, Linguistics of the Tibeto-Burman Area, Natural Language and Linguistic Theory, Phonology, Journal of the International Phonetic Association, Digital Scholarship in the Humanities, Frontiers in Artificial Intelligence

Departmental Service

- New student orientation (IC) organizer, Language Technologies Institute, Carnegie Mellon University, (2021).
- MLT/PhD admissions committee, Language Technologists Institute, Carnegie Mellon University. (2019–2021).
- Faculty Senator, Language Technologies Institute, Carnegie Mellon University (Summer 2020–Spring 2021).
- Curriculum committee, Language Technologies Institute, Carnegie Mellon University. (Fall 2018–Spring 2019)

- MIIS admissions committee, Language Technologies Institute, Carnegie Mellon University. (2018–2019)
- Undergraduate adviser, Department of Linguistics, University of Pittsburgh. (Spring 2008– Spring 2014).
- Colloquium committee, Department of Linguistics, University of Pittsburgh (Fall 2007–2009).
- Curriculum committee, Department of Linguistics, University of Pittsburgh (2007)

Other Professional Activities

- Area Chair, Phonology, Morphology, and Word Segmentation, ACL 2020.
- Area Chair, Phonology, Morphology, and Word Segmentation, EMNLP 2020.
- Program committee, North American Computational Linguistics Olympiad, (2008–2010, 2013–2020).
- Organizing committee, Twenty-Ninth and Thirtieth Annual Meetings of the Berkeley Linguistics Society (2002 and 2003).
- Special session organizer, Twenty-Eighth Annual Meeting of the Berkeley Linguistics Society. 2001. Co-organized with Patrick Chew.