國立臺灣大學文學院語言學研究所
碩士論文

Graduate Institute of Linguistics
College of Liberal Arts

National Taiwan University
Master Thesis

論文題目
Thesis Title

陳蓓怡
Pei-Yi Chen

指導教授：謝舒凱博士
Advisor: Shu-Kai Hsieh, Ph.D.

October 2020
中華民國 109 年 10 月

Abstract

摘要

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Language is constantly changing and evolving. The emergence of new senses, the demise of old ones, and the polysemous nature of lexical items make the process of semantic change a dynamic phenomenon (Robert 2008). As individuals learn new words and meanings throughout their life, so does a language. While recent studies have used time-sliced collections of texts to observe swift meaning changes, the digitalization of texts from earlier time periods opens up research opportunities that incorporates a corpus-driven approach to trace the diachronic development of words and their meanings (Camacho-Collados and Pilehvar 2018; Kutuzov, Øvrelid, et al. 2018; Tahmasebi et al. 2018).

Renouf (2002) reflects on how textual data starts to be treated more than "a static entity." In 1982, Sinclair envisions the possibility of "vast, slowing changing stores of text" and "detailed evidence of language evolution" (as cited in Renouf 2002). The use of digitalized libraries as rich linguistic resources to observe how certain linguistic features are "assimilated" into the language becomes more and more feasible (Renouf 2002).

Additionally, the change in meaning is captured by translating discrete linguistic data into numeric vectors such as word embeddings, especially after the release of Word2vec (Mikolov et al. 2013), GloVe (Pennington et al. 2014) and FastText (Bojanowski et al. 2016). An initial attempt is to generate word embeddings from different time spans and explore whether semantic change occurs based on the neighboring words of the target word from each time period.

The concept of home is an ancient, seemingly familiar and encompassing, but tangible one. Various humanities disciplines have sought to grasp the full picture. Defined by the Oxford English Dictionary (OED), the word *home* is "the place where a person or animal dwells" ("Home", 2020). As one of the earliest 1% entries to be included in the OED, this word has 35 main senses and 214 total senses—Home is a physical space, a place where we feel a "sense of belonging [and] comfort", and even a person's "country or native land." In Mandarin Chinese, the MOE Revised Mandarin Chinese Dictionary defines its translated equivalent 家 *jiā* 'home' as "a place where family members live together (眷屬共同生活的場所)", "a private property (私有財產)", and "people in certain professional fields (經營某種行業或具有某種身份的人)" ("Jia", 2015). Yet, how is the concept of home encoded linguistically? Specifically, how is diachrony interacts with synchrony and variations?

Chapter 2

Literature Review

Chapter 3

Methodology

Chapter 4

Results

Chapter 5

Discussion

Chapter 6

Conclusions

Bibliography

Antoniak, Maria and David Mimno (2018). "Evaluating the stability of embedding-based word similarities." In: *Transactions of the Association for Computational Linguistics* 6, pp. 107–119.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski (2014). "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors." In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 238–247.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo (2018). "Quanteda: An R package for the quantitative analysis of textual data." In: *Journal of Open Source Software* 3.30, p. 774.

Bloomfield, Leonard (1933). "Semantic change." In: *Language*. Allen & Unwin. Chap. 24, pp. 425–443.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2016). "Enriching word vectors with subword information." In: URL: https://arxiv.org/abs/1607.04606.

Brezina, Vaclav (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press.

Camacho-Collados, Jose and Mohammad Taher Pilehvar (2018). "From word to sense embeddings: A survey on vector representations of meaning." In: *Journal of Artificial Intelligence Research* 63, pp. 743–788.

Chen, Keh-Jiann, Chu-Ren Huang, Li-Ping Chang, and Hui-Li Hsu (1996). "Sinica Corpus: Design methodology for balanced corpora." In: *Language*, pp. 167–176.

Chen, Meng-Ying and Zhao-Qing Fu 沈孟穎, 傅朝卿 (2015). "Transformation of modern residential design in Taiwan: A case study on public housing projects from 1920s to 1960s." 台灣現代住宅設計之轉化: 以 1920 年代至 1960 年代公共 (國民) 住宅為例. In: *Journal of Design.* 設計學報 20.4, pp. 43–62.

Coenen, Andy, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg (2019). "Visualizing and measuring the geometry of BERT." In: *Advances in Neural Information Processing Systems*, pp. 8594–8603.

Danescu-Niculescu-Mizil, Cristian, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts (2013). "No country for old members: User lifecycle and linguistic change in online communities." In: *Proceedings of the 22nd International Conference on World Wide Web*, pp. 307–318.

Davies, Mark (2012). "Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English." In: *Corpora* 7.2, pp. 121–157.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding." In: URL: https://arxiv.org/abs/1810.04805.

Dubossarsky, Haim (2018). "Semantic change at large: A computational approach for semantic change research." PhD thesis. Hebrew University of Jerusalem.

Dubossarsky, Haim, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg (2019). "Time-Out: Temporal referencing for robust modeling of lexical semantic change." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 457–470.

Dubossarsky, Haim, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman (2015). "A bottom up approach to category mapping and meaning change." In: *Proceedings of the NetWordS Final Conference*, pp. 66–70.

Dubossarsky, Haim, Daphna Weinshall, and Eitan Grossman (2017). "Outta control: Laws of semantic change and inherent biases in word representation models." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1136–1145.

Ethayarajh, Kawin (2019). "How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65.

Firth, John Rupert (1957). *Modes of meaning, papers in linguistics, 1934-1951*. Oxford University Press.

Fortson IV, Benjamin W (2017). "An approach to semantic change." In: *The Handbook of Historical Linguistics*, pp. 648–666.

Gablasova, Dana, Vaclav Brezina, and Tony McEnery (2017). "Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence." In: *Language learning* 67.S1, pp. 155–179.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou (2018). "Word embeddings quantify 100 years of gender and ethnic stereotypes." In.

Geeraerts, Dirk (1997). *Diachronic prototype semantics: A contribution to historical lexicology*. Oxford University Press.

Giulianelli, Mario (2019). "Lexical semantic change analysis with contextualised word representations." MA thesis. University of Amsterdam.

Goldberg, Yoav and Omer Levy (2014). "Word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method." In: URL: https://arxiv.org/pdf/1402.3722.

Gries, Stefan Th and Martin Hilpert (2012). "Variability-based neighbor clustering: A bottom-up approach to periodization in historical linguistics." In: *The Oxford Handbook of the History of English*, pp. 134–144.

Hales, Alfred W and Robert I Jewett (2009). "Regularity and positional games." In: *Classic Papers in Combinatorics*. Springer, pp. 320–327.

Hamilton, William L, Jure Leskovec, and Dan Jurafsky (2016a). "Cultural shift or linguistic drift? Comparing two computational measures of semantic change." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*. NIH Public Access, pp. 2116–2121.

Hamilton, William L, Jure Leskovec, and Dan Jurafsky (2016b). "Diachronic word embeddings reveal statistical laws of semantic change." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pp. 1489–1501.

Hellrich, Johannes and Udo Hahn (2017a). "Don't get fooled by word embeddings-Better watch their neighborhood." In: *Digital Humanities*, pp. 250–252.

Hellrich, Johannes and Udo Hahn (2017b). "Exploring diachronic lexical semantics with JeSemE." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 31–36.

Hilpert, Martin (2019). "Historical linguistics." In: *Cognitive Linguistics-A Survey of Linguistic Subfields*, pp. 108–131.

Hilpert, Martin and Stefan Th Gries (2009). "Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition." In: *Literary and Linguistic Computing* 24.4, pp. 385–401.

Home (2020). In: *The Oxford English Dictionary*. Last accessed: 2020-09-03. URL: https://www.oed.com/view/Entry/87869?rskey=OqFwzy&result=1#contentWrapper.

Hu, Renfen, Shen Li, and Shichen Liang (2019). "Diachronic sense modeling with deep contextualized word embeddings: An ecological view." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3899–3908.

Huang, Hen-Hsen, Chuen-Tsai Sun, and Hsin-Hsi Chen (2010). "Classical Chinese sentence segmentation." In: *CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pp. 15–22.

Huang, Xiaolei and J. Michael Paul (2019). "Neural temporality adaptation for document classification: Diachronic word embeddings and domain adaptation models." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4113–4123.

Jatowt, Adam, Ricardo Campos, Sourav S Bhowmick, Nina Tahmasebi, and Antoine Doucet (2018). "Every word has its history: Interactive exploration and visualization of word sense evolution." In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 1899–1902. URL: https://doi.org/10.1145/3269206.3269218.

Jia (2015). In: *The MOE Revised Mandarin Chinese Dictionary*. URL: http://dict.revised.moe.edu.tw/cgi-bin/cbdic/gsweb.cgi?o=dcbdic&searchid=W00000005502.

Katricheva, Nadezda, Alyaxey Yaskevich, Anastasiya Lisitsina, Tamara Zhordaniya, Andrey Kutuzov, and Elizaveta Kuzmenko (2020). "Vec2graph: A Python library for visualizing word embeddings as graphs." In: *Analysis of Images, Social Networks and Texts*, pp. 190–198.

Kenter, Tom, Melvin Wevers, Pim Huijnen, and Maarten De Rijke (2015). "Ad hoc monitoring of vocabulary shifts over time." In: *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pp. 1191–1200.

Kutuzov, Andrey and Mario Giulianelli (2020). "UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection." In: URL: https://arxiv.org/abs/2005.00050.

Kutuzov, Andrey, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal (2018). "Diachronic word embeddings and semantic shifts: A survey." In: *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pp. 1384–1397.

Kutuzov, Andrey, Erik Velldal, and Lilja Øvrelid (2017). "Tracing armed conflicts with diachronic word embedding models." In: *Proceedings of the Events and Stories in the News Workshop*, pp. 31–36.

Li, Bai (2020). "Evolution of part-of-speech in Classical Chinese." In: *arXiv preprint arXiv:2009.11144*.

Li, Jiwei, Xinlei Chen, Eduard Hovy, and Dan Jurafsky (2016). "Visualizing and understanding neural models in NLP." In: *Proceedings of NAACL-HLT*, pp. 681–691.

Li, Shen, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du (2018). "Analogical reasoning on Chinese morphological and semantic eelations." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 138–143.

Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila (2016). "Significance testing of word frequencies in corpora." In: *Literary and Linguistic Computing* 31.2, pp. 374–397.

Liu, Chao-Lin and Yi Chang (2019). "Classical Chinese sentence segmentation for tomb biographies of Tang dynasty." In: URL: https://arxiv.org/abs/1908.10606.

Liu, Shusen, Peer-Timo Bremer, Jayaraman J Thiagarajan, Vivek Srikumar, Bei Wang, Yarden Livnat, and Valerio Pascucci (2018). "Visual exploration of semantic relationships in neural word embeddings." In: *IEEE transactions on visualization and computer graphics* 24.1, pp. 553–562.

Mair, Christian (1998). "Corpora and the study of the major varieties of English: Issues and results." In: *The major varieties of English: Papers from MAVEN 97*, pp. 139–158.

Mallett, Shelley (2004). "Understanding home: A critical review of the literature." In: *The sociological review* 52.1, pp. 62–89.

Martinc, Matej, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova (2020). "Capturing evolution in word usage: Just add more clusters?" In: *Companion Proceedings of the Web Conference 2020*, pp. 343–349.

Martinc, Matej, Petra Kralj Novak, and Senja Pollak (2020). "Leveraging contextual embeddings for detecting diachronic semantic shift." In: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pp. 4811–4819.

McCarthy, Diana, Rob Koeling, Julie Weeds, and John A Carroll (2004). "Finding predominant word senses in untagged text." In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistic*, pp. 279–286.

Meng, Yuxian, Xiaoya Li, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li (2019). "Is word segmentation necessary for deep learning of Chinese representations?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pp. 3242–3252.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient estimation of word representations in vector space." In: URL: https://arxiv.org/abs/1301.3781.

Miller, George A. and Walter G. Charles (1991/2007). "Contextual correlates of semantic similarity." In: *Language and Cognitive Processes* 6.1, pp. 1–28.

Moore, Jeanne (2000). "Placing home in context." In: *Journal of environmental psychology* 20.3, pp. 207–217.

Moss, Adam (2020). "Detecting lexical semantic change using probabilistic Gaussian word embeddings." MA thesis. Department of Linguistics and Philology, Uppsala University. URL: https://arxiv.org/pdf/2007.16006.pdf.

Nerlich, Brigitte and David D. Clarke (2001). "Serial metonymy: A study of reference-based polysemisation." In: *Journal of Historical Pragmatics* 2.2, pp. 245–272.

Nielsen, Finn Årup and Lars Kai Hansen (2020). "Creating semantic representations." In: *Statistical Semantics*. Springer, pp. 11–31.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning (2014). "Glove: Global vectors for word representation." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Renouf, Antoinette (2002). "The time dimension in modern English corpus linguistics." In: *Teaching and learning by doing corpus analysis*. Brill Rodopi, pp. 27–41.

Robert, Stéphane (2008). "Words and their meanings: Principles of variation and stabilization." In: *From polysemy to semantic change: Towards a typology of lexical semantic associations*. Ed. by Martine Vanhove. Vol. 106. John Benjamins, pp. 55–92.

Rohrdantz, Christian, Annette Hautli, Thomas Mayer, Miriam Butt, Daniel A. Keim, and Frans Plank (2011). "Towards tracking semantic change via visual analytics." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pp. 305–310.

Rosenfeld, Alex and Katrin Erk (2018). "Deep neural models of semantic shift." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 474–484.

Rychlỳ, Pavel (2008). "A lexicographer-friendly association score." In: *Proceedings of Recent Advances in Slavonic Natural Language Processing (RASLAN)*, pp. 6–9.

Sagi, Eyal, Stefan Kaufmann, and Brady Clark (2011). "Tracing semantic change with Latent Semantic Analysis." In: *Current Methods in Historical Semantics* 73, pp. 161–183.

Samanani, Farhan and Johannes Lenhard (2019). "House and home." In: *The Cambridge Encyclopedia of Anthropology*. Ed. by Felix Stein, Sian Lazar, Matei Candea, Hildegard Diemberger, Joel Robbins, Andrew Sanchez, and Rupert Stasch. URL: http://doi.org/10.29164/19home.

Schlechtweg, Dominik, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi (2020). "SemEval-2020 Task 1: Unsupervised lexical semantic change detection." In: *Proceedings of the 14th International Workshop on Semantic Evaluation*.

Schlechtweg, Dominik, Sabine Schulte im Walde, and Stefanie Eckmann (2018). "Diachronic Usage Relatedness (DURel): A framework for the annotation of lexical semantic change." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 169–174.

Siirtola, Harri, Terttu Nevalainen, Tanja Säily, and Kari-Jouko Räihä (2011). "Visualisation of text corpora: A case study of the PCEEC." In: *How to deal with data: Problems and approaches to the investigation of the English language over time and space*.

Sinclair, John (1982). "Reflections on computer corpora in English language research." In: *Computer corpora in English language research*, pp. 1–6.

Sixsmith, Judith (1986). "The meaning of home: An exploratory study of environmental experience." In: *Journal of environmental psychology* 6.4, pp. 281–298.

Smetanin, Sergey (2018). *Google News and Leo Tolstoy: Visualizing Word2Vec word embeddings using t-SNE*. URL: https://towardsdatascience.com/google-news-and-leo-tolstoy-visualizing-word2vec-word-embeddings-with-t-sne-11558d8bd4d.

Smilkov, Daniel, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B Viégas, and Martin Wattenberg (2016). "Embedding Projector: Interactive visualization and interpretation of embeddings." In: URL: https://arxiv.org/pdf/1611.05469v1.pdf.

Sturgeon, Donald (2019). "Chinese Text Project: A dynamic digital library of premodern Chinese." In: *Digital Scholarship in the Humanities*.

Sturgeon, Donald (2020). *Chinese Text Project*. https://ctext.org. Last accessed: 2020-09-03.

Szymanski, Terrence (2017). "Temporal word analogies: Identifying lexical replacement with diachronic word embeddings." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 448–453.

Tahmasebi, Nina, Lars Borin, and Adam Jatowt (2018). "Survey of computational approaches to diachronic conceptual change." In: URL: https://arxiv.org/abs/1811.06278.

Tang, Xuri (2018). "A state-of-the-art of semantic change computation." In: *Natural Language Engineering* 24.5, pp. 649–676.

Traugott, Elizabeth Closs and Richard B Dasher (2001). *Regularity in semantic change*. Cambridge University Press.

Van der Maaten, Laurens and Geoffrey Hinton (2008). "Visualizing data using t-SNE." In: *Journal of Machine Learning Research* 9, pp. 2579–2605.

Vanhove, Martine, ed. (2008). *From polysemy to semantic change: Towards a typology of lexical semantic associations*. Vol. 106. John Benjamins.

Wang, Yun-lu and Gou, Ying [王雲路、郭穎] (2005). "Shì-shuō gǔ-hàn-yǔ zhòng de cí-zhuì "jiā" [試說古漢語中的詞綴 "家" On the suffix "jia" in early Mandarin Chinese]." In: *Gǔ-hàn-yǔ yán-jiù [古汉语研究 Studies on the Ancient Chinese]* 1, pp. 29–33.

Wei, Pei-chuan, P. M. Thompson, Cheng-hui Liu, Chu-Ren Huang, and Chaofen Sun 魏培泉, 譚樸森, 劉承慧, 黃居仁, 孫朝奮 (1997). "Historical corpora for synchronic and diachronic linguistics studies." 建構一個以共時與歷時語言研究為導向的歷史語料庫. In: *Computational Linguistics and Chinese Language Processing* 2.1, pp. 131–145.

Wevers, Melvin and Marijn Koolen (2020). "Digital begriffsgeschichte: Tracing semantic change using word embeddings." In: *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, pp. 1–18.

Wijaya, Derry Tanti and Reyyan Yeniterzi (2011). "Understanding semantic change of words over centuries." In: *Proceedings of the 2011 International Workshop on Detecting and Exploiting Cultural Diversity on the Social Web*, pp. 35–40.

Xu, Yang and Charles Kemp (2015). "A computational evaluation of two laws of semantic change." In: *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci 2015)*, pp. 2703–2708.

Zellig, Harris (1954/2015). "Distributional structure." In: *Word* 10.2-3, pp. 146–162.

Zhang, Xiao-ping [張小平] (2008). *Dāng-dài hàn-yǔ cí-huì fā-zhǎn yán-jiù [*当代汉语词汇发展变化研究 *Studies on Chinese lexicon development in contemporary time]*. Qí-lǔ shū-shè [齐鲁书社 Qilu Press].

Zhou, Jun-xun [周俊勋] (2009). *Zhōng-gǔ hàn-yǔ cí-huì yán-jiù gāng-yào [*中古汉语词汇研究纲要 *Outline of pre-modern Mandarin Chinese lexicon]*. Bā-shǔ shū-shè [巴蜀书社 Ba-shu Press].

Appendix A

Title of Appendix A

Appendix B

Title of Appendix B