SEMINAR IN WRITING SYSTEMS

FALL 2020

CUNY GRADUATE CENTER

Instructor: Prof. Kyle Gorman
Lecture: TBD
Office hours: TBD

SYNOPSIS

This class will tackle two questions: what is writing, and how does it encode language? The first half of the class will consist of lectures on the definition, origins, and typology of writing systems. The second half of the class will be a student-led seminar on topics in writing systems with a focus on text normalization, the decipherment of lost scripts, orthographic reform, and the psycholinguistics of literacy. (We are not discussing the sociolinguistics of writing because that is roughly the topic of a separate seminar being offered in Fall 2020.) Since (as I will argue in the first portion of the class), writing encodes language primarily by means of (morpho)phonological analysis, students should have completed graduate coursework in phonology.

LEARNING GOALS

Students will:

1. learn the history of writing,
2. engage with a linguistically-informed definition of writing,
3. learn to identify key typological features of writing system,
4. become familiar with research questions in text normalization, decipherment, orthographic reform, and the psycholinguisics of literacy.

ACCOMMODATIONS

The instructor will attempt to provide all reasonable accommodations to students upon request. If you believe you are covered under the Americans With Disabilities Act, please direct accommodations requests to Matthew G. Schoengood, Vice President for Student Affairs.

ATTENDANCE

Students are expected to attend all lectures, and as much as 50% of the final grade may reflect attendance and participation in class. The instructor is not responsible for reviewing materials missed due to absence.

GRADING

During the first half of the class, students will be required to complete small assignments or reflections on readings and lectures. During the second half of the class they will be assigned to lead discussions and presentations. At the end of the class, they will submit a research paper on writing systems.

INTEGRITY

In line with the Student Handbook policies on plagiarism, students are expected to complete their own work. The instructor reserves the right to refer violations of this policy to the Academic Integrity Officer.

RESPECT

For the sake of the privacy, students are asked not to record lectures. Students are expected to be considerate of your peers and to treat them with respect during class discussions.

SCHEDULE

(Dates TBD.)

1. ### LECTURES
    1. The ancient Near East
    2. Neo-Semitic alphabets and alphasyllabaries
    3. The ancient Far East
    4. Japanese & Korean
    5. Mayan
    6. Writing system typology
2. ### SEMINARS
    1. Decipherment of ancient scripts
    2. Decipherment of ciphers
    3. Text normalization for speech technologies
    4. Text normalization for user-generated text
    5. Orthographic reform
    6. Psycholinguistics of literacy

TEXTS

The primary text we will use is Rogers 2005, which is worth owning. Other high-quality general-purpose texts on this topic I recommend are Daniels & Bright 1996 (an enormous compendium of surveys: the first place to look to learn about a particular writing system), Gnanadesikan 2008, and Robinson 2007 (both textbooks, but written for undergraduates). However, we will be reading from a wide variety of other resources; this will be a reading-heavy class.

BIBLIOGRAPHY

Barber, E. 1974. *Archaeological decipherment: a handbook*. Princeton: Princeton University Press.

Brody, S., and Diakopoulos, N. 2011. Cooooooooooooooollllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 562-570.

Chadwick, J. 1958. *The decipherment of Linear B*. Cambridge: Cambridge University Press.

Chomsky, N. and Halle, M. 1968. *Sound pattern of English*. New York: Harper & Row.

Chua, M. van Esch, D., Coccaro, N. Cho, E., Bhandari, S., and Jia, L. 2018. Text normalization infrastructure that scales to hundreds of language varieties. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 1353-1356.

Coe, M. 1992. *Breaking the Maya code*. London: Thames & Hudson.

Daniels, P. T. and Bright, William (eds.). 1996. *The world's writing systems*. New York: Oxford University Press.

DeFrancis, J. 1989. *Visible speech: the diverse oneness of writing systems*. Honolulu: University of Hawaii Press.

Ebden, P. and Sproat, R. 2015. The Kestrel TTS text normalization system. *Natural Language Engineering* 21(3): 333-353.

Eisenstein, J. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359-369.

van Esch, D., Chua, M., and Rao, K. (2016). Predicting pronunciations with syllabification and stress with recurrent neural networks. In *INTERSPEECH*, pages 2841-2845.

van Esch, D., Sarbar, E., Lucassen, T., O'Brien, J., Breiner, T., Prasad, M., Crew, E., Nguyen, C. and Beaufays, F. 2019. Writing across the world's languages: deep internationalization for Gboard, the Google keyboard. `arXiv:1912.01218`.

Farmer, S., Sproat, R., and Witzel, M. 2004. The collapse of the Indus-script thesis: the myth of a literate Harappan civilization. *Electronic Journal of Vedic Studies* 11(2): 19-58.

Fox, M. 2014. *The riddle of the labyrinth: the quest to crack an ancient code*. New York: HarperCollins.

Gelb, I. J. 1952. *A study of writing*. Chicago: University of Chicago Press.

Gillick, D. 2009. Sentence boundary detection and the problem with the U.S. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 241-244.

Gnanadesikan, A. E. 2008. *The writing revolution: cuneiform to the internet*. Malden, MA: Wiley-Blackwell.

Gorman, K., and Mazovetskiy, G., and Nikolaev, V. 2018. Improving homograph disambiguation with machine learning. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 1349-1352.

Gorman, K. and Sproat, R. 2016. Minimally supervised number normalization. *Transactions of the Association for Computational Linguistics* 4: 507-519.

Gorman, K. 2018. Another pseudo-decipherment of the Voynich manuscript. Blog post accessed April 17, 2020.

Graehl, J. and Knight, K. 1998. Machine transliteration. *Computational Linguistics* 24(4): 599-612.

Guy, J. 2006. General properties of the Rongorongo writing. *Rapa Nui Journal* 20(1): 53-66.

Han, B. and Baldwin, T. 2011. Lexical normalisation of short text messages: makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368-378.

Harris, William V. 1989. *Ancient literacy*. Cambridge: Harvard University Press.

Hauer, B. and Kondrak, G. 2016. Decoding anagrammed texts written in an unknown language and script. *Transactions of the Association for Computational Linguistics* 4: 75-86.

Hooker, J. T. 1980. *Linear B: an introduction*. Bristol: Bristol Classical Press.

Kessler, B. and Treiman, R. 2003. Is English spelling chaotic? Misconceptions concerning its irregularity. *Reading Psychology* 24(3-4): 267-289.

Knight, K. and Yamada, K. 1999. A computational approach to deciphering unknown scripts. In *Proceeding of the ACL Workshop on Unsupervised Learning in Natural Language Processing*, pages 37-44.

Knight, K., Nair, A., Rathod, N., and Yamada, K. Unsupervised analysis for decipherment problems. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 499-506.

Knight, K., Megyesi, B., and Schaefer, C. 2012. The secrets of the Copiale Cipher. *Journal for Research into Freemasonry and Fraternalism* 2(2): 314-324.

Lee, J. L., Ashby, L. F.E.. Garza, M. E., Lee-Sikka, Y., Miller, S., Wong, A.. McCarthy, A. and Gorman, K. 2020. Massively multilingual pronunciation mining with WikiPron. In *LREC*, to appear.

Lee, R. Jonathan, and Ziman, P. 2010. Pictish symbols revealed as a written language through application of Shannon entropy. *Proceedings of the Royal Society* 466: 2545-2560.

Liu, F., Weng, F., Wang, B., and Liu, Y. 2011. Insertion, deletion, or substitution?: normalizing text messages without pre-categorization nor supervision. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 71-76.

Melena, J. L. 2014. Mycenaean writing. In Duhoux, Y., and Davies, A. M. (ed.), *A companion to Linear B: Mycenaean Greek texts and their world*, pages 1-186. Louvain-la-Neuve, Belgium: Peeters.

Merhav, Y. and Ash, S. 2018. Design challenges in named entity transliteration. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 630-640.

Ng., Axel H., Gorman, K., and Sproat, R. 2017. Minimally supervised written-to-spoken text normalization. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 665-670.

Novak, J. R., Minematsu, N., and Hirose, K. (2016). Phonetisaurus: exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework. *Natural Language Engineering* 22(6): 907-938.

Pope, M. 1999. *The story of decipherment: from Egyptian hieroglyphs to Maya script*. New York: Thames & Hudson.

Rao, R., Yadav, N., Vahia, M., Joglekar, H., Adhikari, R, and Mahadevan, I. 2009. Entropic evidence for linguistic structure in the Indus script. *Science* 342(5931): 1165.

Ravi, S. and Knight, K. 2011. Bayesian inference for Zodiac and other homophonic ciphers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 239-247.

Read. J. 2005. Using emoticons to reduce dependency on machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, pages 43-48.

Reddy, S. and Knight, K. 2011. What we know about the Voynich manuscript. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 78-86.

Roark, B., and Sproat, R. Hippocratic abbreviation expansion. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 364-369.

Robinson, A. 2002. *Lost languages: the enigma of the world's undeciphered scripts*. New York: McGraw-Hill.

Robinson, A. 2007. *The story of writing: alphabets, hieroglyphs & pictograms*. London: Thames & Hudson.

Robinson, A. 2012. *The man who deciphered Linear B: the story of Michael Ventris*. London: Thames & Hudson.

Ritchie, S., Sproat, R., Gorman, K., van Esch, D., Schallhart, C. Bampounis, N., Brard, B., Mortensen, J. F., Holt, M., and Mahon, E. 2019. Unified verbalization for speech recognition & synthesis across languages. In *INTERSPEECH*, pages 3530-3534.

Rogers, Henry. 2005. *Writing systems: a linguistic approach*. Cambridge: Blackwell.

Rugg, G. and Taylor, G. 2017. Hoaxing statistical features of the Voynich Manuscript. *Cryptologia* 41(3): 247-268.

Scribner, S. and Cole, M. 1981. *The psychology of literacy*. Cambridge: Harvard University Press.

Sproat, R. 2000. *A computational theory of writing systems*. Cambridge: Cambridge University Press.

Sproat, R., Black, A. W., Chen, S., Kumar, S., Ostendorf, M., and Richards, C. 2001. Normalization of non-standard words. *Computer Speech & Language* 15(3): 287-333.

Sproat, R. 2010a. Lightly supervised learning of text normalization: Russian number names. In *IEEE Workshop on Speech and Language Technology*, pages 436-441.

Sproat, R. 2010b. *Language, technology, and society*. Oxford: Oxford University Press.

Sproat, R. 2010c. Ancient symbols, computational linguistics, and the reviewing practices of the general science journals. *Computational Linguistics* 36(3): 585-594.

Sproat, R. 2014. A statistical comparison of written language and nonlinguistic symbol systems. *Language* 90(2): 457-481.

Sproat, R. and Hall, K. 2014. Applications of maximum entropy rankers to problems in spoken language processing. In *INTERSPEECH*, pages 761-764.

Taylor, P. 2005. Hidden Markov models for grapheme to phoneme conversion. In *INTERSPEECH*, pages 1973-1976.

Zhang, H., Sproat, R, Ng, A. H., Stahlberg, F., Peng, X., Gorman, K., and Roark, B. 2019. Neural models of text normalization for speech applications. *Computational Linguistics* 45(2): 293-337.