Reading list for SPR4104

Diana Santos

Spring 2016, modified in 2020

Here I list, first, the textbooks from which I take examples or ideas, and then I give the full reference of the articles that I ask the students to read (and which constitute the reading list proper).

1 Textbooks

In addition to the obligatory Baayen (2008), I use as inspiration or source of data Johnson (2008) and Gries (2009), which are also books on statistics for linguists using R; while I often cite Crawley (2005) and Fox e Weisberg (2011) for clear explanations of statistics with R, although not on linguistic topics.

Other books that I use, but which neither deal with R, nor with linguistics in particular, are Devore e Berk (1995), Simon (1999), Agresti (1996) and Cohen (1995).

For Norwegian terms I used Lund e Christophersen (1999), Hagen (2014) and Aalen (1999).

For distributional semantics, I use Widdows (2004), Dorow (2006) and Mihalcea e Radev (2011).

Other articles or books cited in the lectures are Clark (1973), Dunning (1993) and Kenny (1982) and Oakes (2014). I also use the material from my own ESSLLI summer school courses, Santos (2007) and Santos e Finatto (2010).

Newer books also of interest are Levshina (2015) and Brezina (2018), as well as Jockers (2014) for statistical analysis of literature.

2 Reading list

These are the articles that the students are supposed to read during the term:

1. Karlgren, Hans. "Quantitative models – of what?", *Statistical Methods in Linguistics, SMIL*, 1975, pp. 25-31.

2. Church, Kenneth W. "Empirical estimates of adaptation: the chance of two noriegas is closer to p/2 than p2." In *Proceedings of the 18th Conference on Computational Linguistics* - Volume 1 (Saarbrücken, Germany, July 31 - August 04, 2000). International Conference On Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, pp. 180-186.

3. Schmidt, Kari Anne Rand. "Male and female language in Jane Austen's novels". In Stig Johansson & Bjørn Tysdahl (eds.), *Papers from the First Nordic Conference for English Studies (Oslo, 19-19 September, 1980)* , pp. 198-210.

4. Biber, Douglas. "Investigating macroscopic textual variation through multifeature/multidimensional analyses". *Linguistics* 23.2, 1985, Mouton publishers, pp. 337-360.

5. Gilqin, Gaëtanelle. "The Integrated Contrastive Model: Spicing up your data", *Languages in Contrast* 3, 1, 2000/2001, pp. 95-123.

6. Koppel, Moshe & Noam Ordan. "Translationese and its dialects". In *Proceedings of ACL 2011*, pp. 1318-1326.

7. Halliday, M.A.K. "Corpus studies and probabilistic grammar". In Aijmer, Karin & Bengt Altenberg (eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, Longman, 1991, pp. 30-43.

8. Grefenstette, Gregory. "Corpus-derived first, second and third-order word affinities". In Willy Martin, Willem Meijs, Margreet Moerland, Elsemiek ten Pas, Piet van Sterkenburg, and Piek Vossen (eds.), *Euralex '94 Proceedings*, Amsterdam: Vrije Universiteit, pp. 279-290.

9. Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhuw. (2001). "BLEU: a Method for Automatic Evaluation of Machine Translation", Research Report, Computer Science IBM Research Division, T.J. Watson Research Center, RC22176 (W0109-022), 17 September 2001, `http://domino.watson.ibm.com/library/CyberDig.nsf/Home`.

10. Ioannidis, John P. "Why most Published Research Findings Are False", *PlosMed* 2, 8, August 2005, pp. 696-701.

11. Weaver, Warren. "Probability, rarity, interest and surprise". *Scientific monthly* 67, 1948, pp. 390-392.

References

Odd O. Aalen. *Innføring i statistikk med medisinske eksempler*, Ad Notam Gyldendal, 1999. 2. utgave.

Alan Agresti. *An Introduction to Categorical Data Analysis*, John Wiley and Sons, 1996.

Harald Baayen. *Analyzing Linguistic Data: A practical introduction to Statistics using R*, Cambridge University Press, 2008.

Vaclav Brezina. *Statistics in Corpus Linguistics: A Practical Guide*, Cambridge University Press, 2018.

Herbert H. Clark. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of verbal learning and verbal behavior*, 12:335–59, 1973.

Paul R. Cohen. *Empirical Methods for Artificial Intelligence*, The MIT Press, 1995.

Michael J. Crawley. *Statistics: An Introduction using R*, John Wiley and Sons, 2005.

Jay L. Devore e Kenneth N. Berk. *Modern Mathematical Statistics with Applications*, Thomson Brooks/Cole, 1995.

Beate Dorow. *A Graph Model for Words and their Meanings*. Tese de doutoramento, IMS, Universidade de Stuttgart, 2006.

Ted Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1):61–74, March de 1993.

John Fox e Sanford Weisberg. *An R companion to applied regression*, Sage, 2n edition, 2011.

Stefan Th. Gries. *Statistics for Linguistics with R: A Practical Introduction*, Mouton de Gruyter, 2009.

Per Christian Hagen. *Innføring i sannsynlighetsregning og statistikk*, Cappelen Damm Akademisk, 2014.

Matthew L. Jockers. *Text analysis with R for Students of Literature*, Springer, 2014.

Keith Johnson. *Quantitative Methods in Linguistics*, Wiley-Blackwell, 2008.

Anthony Kenny. *The computation of style: an introduction to statistics for students of literature and humanities*, Pergamon Press, 1982.

Natalia Levshina. *How to do Linguistics with R: Data exploration and statistical analysis*, John Benjamins, 2015.

Thorleif Lund e Knut-Andreas Christophersen. *Innføring i statistikk*, Universitetsforlag, 1999.

Rada Mihalcea e Dragomir Radev. *Graph-Based Natural Language Processing and Information Retrieval*, Cambridge University Press, 2011.

Michael P. Oakes. *Literary Detective Work on the Computer*, John Benjamins Publishing Co., 2014.

Diana Santos. Evaluation in natural language processing. 6-17 August de 2007. `http://www.linguateca.pt/Diana/download/EvaluationESSLLI07.pdf`.

Diana Santos e Maria José Bocorny Finatto. Words and their secrets. 2010.

Julian L. Simon. *The Philosophy and Practice of Statistics and Resampling*, 1999. `http://www.juliansimon.com/writings/Resampling_Philosophy/`. unfinished.

Dominic Widdows. *Geometry and Meaning*, CSLI Publications, 2004.