Computational study of linguistic differences.

Winter semester 2015

Course Description:

This seminar focuses on computational methods for studying language change and variation. Languages differ based on time, geography, speakers (or author), or even audience. This course is concerned with how to find these differences, how to reason about them, and how to deal with them in computational linguistic methods and applications. In this course, we will read and discuss a selection of articles mainly in the areas of computational study of dialectology (dialectometry) and computational study of literary style, or stylometry, but also in other relevant areas of language change and variation depending on the participants' interests.

Prerequisites: The course requires familiarity with basic computational linguistics and machine learning methods. The course is aimed at students in a masters' program, which assumes serious motivation and academic maturity.

Course work and evaluation:

- Active participation in class (20%).
  All students have to read the papers to be discussed in the class, and expected to participate in the discussion. All students are required to send two questions for each paper to the instructor via email not later than 10:00 on the day the article is discussed in the class.
- Presenting a paper (30%).
- Write and submit a term paper (50%).
  The term paper should ideally be written on a practical application/experimentation relevant to one of the course topics. Team work is possible if agreed with the instructor in advance.

Language:

The course language is English.

Instructor: Çağrı Çöltekin
Office: Room 2.24, Blochbau (Wilhelmstr. 19)
Email: ccoltekin@sfs.uni-tuebingen.de
Office hours: TBA

Extended bibliography

The following is a list of potential papers to be discussed in the class. Students are encouraged to propose other reading material.

1. Efstathios Stamatatos (2009). "A survey of modern authorship attribution methods". In: *Journal of the American Society for information Science and Technology* 60.3, pp. 538–556

2. Walter Daelemans (2013). "Explanation in computational stylometry". In: *Computational Linguistics and Intelligent Text Processing*. Springer, pp. 451–462

3. Hans Goebl (1993). "Dialectometry: a short overview of the principles and practice of quantitative classification of linguistic atlas data". In: ed. by Reinhard Köhler and Burghard B Rieger. Springer Science & Business Media, pp. 277–315

4. Martijn Wieling and John Nerbonne (2015). "Advances in dialectometry". In: *Annu. Rev. Linguist.* 1.1, pp. 243–264

5. Matthew L. Jockers and Daniela M. Witten (2010). "A comparative study of machine learning methods for authorship attribution". In: *Literary and Linguistic Computing* 25.2, pp. 215–223. DOI: 10.1093/llc/fqq001

6. Jack Grieve (2007). "Quantitative Authorship Attribution: An Evaluation of Techniques". In: *Literary and Linguistic Computing* 22.3, pp. 251–270. DOI: 10.1093/llc/fqm020

7. John Burrows (2002). "'Delta': A measure of stylistic difference and a guide to likely authorship". In: *Literary and Linguistic Computing* 17.3, pp. 267–287. DOI: 10.1093/llc/17.3.267

8. John Burrows (2007). "All the way through: testing for authorship in different frequency strata". In: *Literary and Linguistic Computing* 22.1, pp. 27–47. DOI: 10.1093/llc/fqi067

9. David L Hoover (2004). "Testing Burrows's delta". In: *Literary and linguistic computing* 19.4, pp. 453–475

10. Shlomo Argamon (2008). "Interpreting Burrows's Delta: Geometric and probabilistic foundations". In: *Literary and Linguistic Computing* 23.2, pp. 131–147

11. David L Mealand (1995). "Correspondence analysis of Luke". In: *Literary and linguistic computing* 10.3, pp. 171–182

12. Jacques Savoy (2013). "Authorship attribution based on a probabilistic topic model". In: *Information Processing & Management* 49.1, pp. 341–354

13. Jose Nilo G Binongo (2003). "Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution". In: *Chance* 16.2, pp. 9–17

14. Michael Oakes and Alois Pichler (2013). "Computational Stylometry of Wittgenstein's "Diktat für Schlick"." In: *Bergen Language and Linguistics Studies* 3.1

15. Richard S Forsyth and Serge Sharoff (2014). "Document dissimilarity within and across languages: A benchmarking study". In: *Literary and Linguistic Computing* 29.1, pp. 6–22

16. Carmen Klaussner, John Nerbonne, and Çağrı Çöltekin (2015). "Finding Characteristic Features in Stylometric Analysis". In: *Journal of Digital Scholarship in the Humanities* (to appear)

17. Conrad Sanderson and Simon Guenter (2006). "Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation". In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 482–491

18. Fazli Can and Jon M Patton (2004). "Change of writing style with time". In: *Computers and the Humanities* 38.1, pp. 61–82

19. James W Pennebaker and Lori D Stone (2003). "Words of wisdom: language use over the life span." In: *Journal of personality and social psychology* 85.2, p. 291

20. Mats Dahllöf (2012). "Automatic prediction of gender, political affiliation, and age in Swedish politicians from the wording of their speeches—A comparative study of classifiability". In: *Literary and linguistic computing* 27.2, pp. 139–153

21. Peter Garrard, Lisa M Maloney, John R Hodges, and Karalyn Patterson (2005). "The effects of very early Alzheimer's disease on the characteristics of writing by a renowned author". In: *Brain* 128.2, pp. 250–260

22. Xuan Le, Ian Lancashire, Graeme Hirst, and Regina Jokel (2011). "Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists". In: *Literary and Linguistic Computing* 26.4, pp. 435–461

23. Kim Luyckx and Walter Daelemans (1998). "Using syntactic features to predict author personality from text". In: vol. 22, pp. 319–346

24. John Noecker, Michael Ryan, and Patrick Juola (2013). "Psychological profiling through textual analysis". In: *Literary and Linguistic Computing* 28.3, pp. 382–387

25. Heather F Windram, Prue Shaw, Peter Robinson, and Christopher J Howe (2008). "Dante's Monarchia as a test case for the use of phylogenetic methods in stemmatic analysis". In: *Literary and Linguistic Computing* 23.4, pp. 443–463

26. Patrick Juola and R Harald Baayen (2005). "A controlled-corpus experiment in authorship identification by cross-entropy". In: *Literary and Linguistic Computing* 20.Suppl, pp. 59–67

27. Kim Luyckx and Walter Daelemans (2005). "Shallow Text Analysis and Machine Learning for Authorship Attribtion". In: *Selected papers from the Fifteenth CLIN Meeting*. Vol. 4. LOT, pp. 149–160

28. Özlem Uzuner and Boris Katz (2005). "A comparative study of language models for book and author recognition". In: *Natural Language Processing–IJCNLP 2005*. Springer, pp. 969–980

29. Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan (2007). "Stylistic text classification using functional lexical features". In: *Journal of the American Society for Information Science and Technology* 58.6, pp. 802–822

30. Paul Clough (2000). *Plagiarism in natural and programming languages: an overview of current tools and technologies*. Tech. rep. CS-00-05. Department of Computer Science, University of Sheffield, UK

31. Paul Clough and Mark Stevenson (2011). "Developing a corpus of plagiarised short answers". In: *Language Resources and Evaluation* 45.1, pp. 5–24

32. Daniel Bär, Torsten Zesch, and Iryna Gurevych (2012). "Text reuse detection using a composition of text similarity measures". In: *Proceedings of COLING*. vol. 1, pp. 167–184

33. Maxim Mozgovoy, Tuomo Kakkonen, and Erkki Sutinen (2007). "Using natural language parsers in plagiarism detection." In: *Proceedings of Speech and Language Technology in Education (SLaTE 2007)*, pp. 77–79

34. Mario Zechner, Markus Muhr, Roman Kern, and Michael Granitzer (2009). "External and intrinsic plagiarism detection using vector space models". In: *Proc. of 25th Annual conference of the Spanish society for natural language processing (SEPLN 2009)*, pp. 38–46

35. Hugo T. Jankowitz (1988). "Detecting Plagiarism in Student Pascale Programs". In: *The Computer Journal* 31.1, pp. 1–8

36. Enrique Flores, Alberto Barrón-Cedeno, Paolo Rosso, and Lidia Moreno (2011). "Towards the detection of cross-language source code reuse". In: *Natural Language Processing and Information Systems*. LNCS. Springer, pp. 250–253

37. Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov (2010). "Identification of translationese: A machine learning approach". In: *Computational linguistics and intelligent text processing*. Springer, pp. 503–511

38. Ella Rabinovich and Shuly Wintner (2015). "Unsupervised Identification of Translationese". In: *Transactions of the Association for Computational Linguistics* 3, pp. 419–432

39. Mehran Sahami, Susan Dumais, David Heckerman, and Eric Horvitz (1998). "A Bayesian approach to filtering junk e-mail". In: *Learning for Text Categorization: Papers from the 1998 workshop*. Vol. 62, pp. 98–105

40. Robert G Shackleton et al. (2005). "English-American Speech Relationships A Quantitative Approach". In: *Journal of English Linguistics* 33.2, pp. 99–160

41. John Nerbonne (2006). "Identifying linguistic structure in aggregate comparison". In: *Literary and Linguistic Computing* 21.4, pp. 463–475

42. Jack Grieve, Dirk Speelman, and Dirk Geeraerts (2011). "A statistical method for the identification and aggregation of regional linguistic variation". In: *Language Variation and Change* 23.02, pp. 193–221

43. Jack Grieve, Dirk Speelman, and Dirk Geeraerts (2013). "A multivariate spatial analysis of vowel formants in American English". In: *Journal of Linguistic Geography* 1.01, pp. 31–51

44. Jelena Prokić and Tim Van de Cruys (2010). "Exploring dialect phonetic variation using PARAFAC". in: *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*. Association for Computational Linguistics, pp. 46–53

45. Tom Ruette and Dirk Speelman (2014). "Transparent aggregation of variables with Individual Differences Scaling". In: *Literary and Linguistic Computing* 29.1, pp. 89–106

46. Dirk Speelman and Dirk Geeraerts (2008). "The role of concept characteristics in lexical dialectometry". In: *International Journal of Humanities and Arts Computing* 2.1-2, pp. 221–242

47. Tom Ruette, Dirk Speelman, and Dirk Geeraerts (2013). "Lexical variation in aggregate perspective". In: *Pluricentricity: Language Variation and Sociocognitive Dimensions* 24, p. 103

48. Kris Heylen and Tom Ruette (2013). "Degrees of semantic control in measuring aggregated lexical distances". In: *Approaches to Measuring Linguistic Differences, edited by Lars Borin and Anju Saxena*, pp. 353–374

49. Therese Leinonen (2011). "Aggregate analysis of vowel pronunciation in Swedish dialects". In: *Oslo Studies in Language* 3.2

50. Jelena Prokić and Michael Cysouw (2013). "Combining regular sound correspondences and geographic spread". In: *Language Dynamics and Change* 3.2, pp. 147–168

51. John Nerbonne (2010). "Measuring the Diffusion of Linguistic Change". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 365. Special issue papers from "Cultural and Linguistic Diversity", conference held at AHRC Centre for Evolution of Cultural Diversity, London, Dec. 9-13, 2008, pp. 3821–3828. DOI: 10.1098/rstb.2010.0048

52. Marco René Spruit, Wilbert Heeringa, and John Nerbonne (2009). "Associations among linguistic levels". In: *Lingua* 119.11, pp. 1624–1642

53. Jelena Prokić, Çağrı Çöltekin, and John Nerbonne (2012). "Detecting shibboleths". In: *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. Association for Computational Linguistics, pp. 72–80

54. Marco René Spruit, Wilbert Heeringa, and John Nerbonne (2009). "Associations among linguistic levels". In: *Lingua* 119.11, pp. 1624–1642

55. Marco René Spruit, Wilbert Heeringa, and John Nerbonne (2009). "Associations among linguistic levels". In: *Lingua* 119.11, pp. 1624–1642

56. Jack Grieve (2012). "A statistical analysis of regional variation in adverb position in a corpus of written Standard American English". In: *Corpus linguistics and linguistic theory* 8.1, pp. 39–72

57. Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing (2010). "A latent variable model for geographic lexical variation". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1277–1287

58. Jonathan Harrington, Sallyanne Palethorpe, and Catherine I. Watson (2000). "Does the Queen speak the Queen's English?." In: *Nature* 408.6815, p. 927

59. Nairan Ramirez-Esparza, Cindy K Chung, Ewa Kacewicz, and James W Pennebaker (2008). "The Psychology of Word Use in Depression Forums in English and in Spanish: Texting Two Text Analytic Approaches." In: *International Conference on Weblogs and Social Media*, pp. 102–108

Some additional links for relevant conferences/tools:

- PAN workshop series: http://pan.webis.de/

- A recent workshop on closely related langauges: `http://ttg.uni-saarland.de/lt4vardial2015/`.
- Tools for dialectometry:
  - Gabmap `http://www.gabmap.nl`
  - Visual DielectoMetry: `http://ald.sbg.ac.at/dm/germ/VDM/`.
- R package "Stylistics in R"