



Universiteit
Leiden
The Netherlands

Contributions to the computational processing of diachronic linguistic corpora

Cunha, E. Landulfo Teixeira Paradela

Citation

Cunha, E. L. T. P. (2020, March 19). *Contributions to the computational processing of diachronic linguistic corpora*. LOT dissertation series. LOT, Amsterdam. Retrieved from <https://hdl.handle.net/1887/133504>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/133504>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/133504> holds various files of this Leiden University dissertation.

Author: Cunha, E. Landulfo Teixeira Paradela

Title: Contributions to the computational processing of diachronic linguistic corpora

Issue date: 2020-03-19

Bibliography

- Acerbi, A., Lampos, V., Garnett, P., and Bentley, R. A. (2013). The expression of emotions in 20th century books. *PLOS ONE*, 8(3):e59030.
- Algeo, J. and Algeo, A. S., editors (1993). *Fifty years Among the New Words: A dictionary of neologisms, 1941–1991*. Cambridge University Press, Cambridge.
- Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236.
- American Dialect Society (2018). “Fake news” is 2017 American Dialect Society word of the year. Retrieved from <https://bit.ly/2FCfbKX> . Accessed on January 4, 2019.
- Amoris, V., Gollner, A. P., Goulart, E., and Pessoni, A. (2012). Marketing político e redes sociais: Reflexos nas eleições 2010 à Presidência da República. In Queiroz, A. C. F., Tomaziello, P. S., and Macedo, R. G., editors, *Comunicação política e eleitoral no Brasil: Perspectivas e limitações no dinamismo político*, pages 140–158. POLITICOM, Americana.

- Arif, N., Al-Jefri, M., Bizzi, I. H., Perano, G. B., Goldman, M., Haq, I., Chua, K. L., Mengozzi, M., Neunez, M., Smith, H., and Ghezzi, P. (2018). Fake news or weak science? Visibility and characterization of anti-vaccine webpages returned by Google in different languages and countries. *Frontiers in Immunology*, 9:1215.
- Arnaudo, D. (2017). Computational propaganda in Brazil: Social bots during elections. *Computational Propaganda Research Project*, 8.
- Ayto, J. (1989). *The Longman register of new words*, volume 1. Longman, Harlow.
- Ayto, J. (1990). *The Longman register of new words*, volume 2. Longman, Harlow.
- Ayto, J. (1999). *Twentieth century words*. Oxford University Press, Oxford.
- Backus, A. (2014). Towards a usage-based account of language change: Implications of contact linguistics for linguistic theory. In Nicolaï, R., editor, *Questioning language contact: Limits of contact, contact at its limits*, volume 1 of *Brill Studies in Language Contact and Dynamics of Language*, pages 91–118. Brill, Leiden.
- Baker, P. (2013). Will Ms ever be as frequent as Mr? A corpus-based comparison of gendered terms across four diachronic corpora of British English. *Gender and Language*, 4(1):125–149.
- Baker, P., Hardie, A., and McEnery, T. (2006). *A glossary of corpus linguistics*. Edinburgh University Press, Edinburgh.
- Barker, E., Paramita, M., Aker, A., Kurtic, E., Hepple, M., and Gaizauskas, R. (2016). The SENSEI Annotated Corpus: Human summaries of reader comment conversations in on-line news. In

Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2016), pages 42–52. Association for Computational Linguistics (ACL).

Baron, N. (1994). Computer languages. In Asher, R. and Simpson, J., editors, *The encyclopedia of language and linguistics*, volume 2, pages 662–670. Pergamon Press, Oxford.

Bauer, L. (2002). Inferring variation and change from public corpora. In Chambers, J., Trudgill, P., and Schilling-Estes, N., editors, *The handbook of language variation and change*, pages 97–114. Wiley Blackwell, Hoboken.

BBC News (2018). Fake news: French language body urges alternative phrase. BBC News. Retrieved from <https://www.bbc.com/news/world-europe-45754756> . Accessed on January 4, 2019.

Berber Sardinha, T. (2013). Variação entre registros da Internet. In Shepherd, T. G. and Saliés, T. G., editors, *Linguística da internet*, pages 55–75. Contexto, São Paulo.

Berger, G. (2009). How the Internet impacts on international news: Exploring paradoxes of the most global medium in a time of ‘hyperlocalism’. *International Communication Gazette*, 71(5):355–371.

Biber, D. and Finegan, E. (2001). Diachronic relations among speech-based and written registers in english. In Conrad, S. and Biber, D., editors, *Variation in English: Multi-dimensional studies*, pages 66–83. Routledge, Oxon/New York.

Biber, D. and Gray, B. (2011). Grammatical change in the noun phrase: The influence of written language use. *English Language and Linguistics*, 15(2):223–250.

- Bird, S., Loper, E., and Klein, E. (2009). *Natural language processing with Python*. O'Reilly Media Inc.
- Bird, S. and Simons, G. (2003). Seven dimensions of portability for language documentation and description. *Language*, 79(3):557–582.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- Blumenthal-Dramé, A. (2012). *Entrenchment in usage-based theories: What corpus data do and do not reveal about the mind*, volume 83 of *Topics in English Linguistics*. De Gruyter Mouton, Berlin.
- Bochkarev, V., Solovyev, V., and Wichmann, S. (2014). Universals versus historical contingencies in lexical evolution. *Journal of the Royal Society Interface*, 11.
- Bowker, L. and Pearson, J. (2002). *Working with specialized language: A practical guide to using corpora*. Routledge, London.
- Bradshaw, S. and Howard, P. N. (2018). Challenging truth and trust: A global inventory of organized social media manipulation. *The Computational Propaganda Research Project*.
- Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology*, 1(1):44–59.
- Brummette, J., DiStaso, M., Vafeiadis, M., and Messner, M. (2018). Read all about it: The politicization of “fake news” on Twitter. *Journalism & Mass Communication Quarterly*, 95(2):497–517.
- Buechel, S., Hellrich, J., and Hahn, U. (2016). Feelings from the past – Adapting affective lexicons for historical emotion analysis. In

Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), pages 54–61.

- Burger, P., Kanhai, S., Pleijter, A., and Verberne, S. (2019). The reach of commercially motivated junk news on Facebook. *PLOS ONE*, 14(8):e0220446.
- Bussmann, H. (1996). *Routledge dictionary of language and linguistics*. Routledge, New York.
- Byrd, R. J., Calzolari, N., Chodorow, M. S., Klavans, J. L., Neff, M. S., and Rizk, O. A. (1987). Tools and methods for computational lexicology. *Computational Linguistics*, 13(3-4):219–240.
- Bârgăoanu, A. and Radu, L. (2018). Fake news or disinformation 2.0? Some insights into Romanians’ digital behaviour. *Romanian Journal of European Affairs*, 18(1):24–38.
- Caetano, J. A., Magno, G., Cunha, E. L. T. P., Meira Jr., W., Marques-Neto, H. T., and Almeida, V. (2018). Characterizing the public perception of WhatsApp through the lens of media. In *Proceedings of the 2nd International Workshop on Rumours and Deception in Social Media (RDSM 2018)*.
- Caetano, J. A., Magno, G., Gonçalves, M. A., Almeida, J., Marques-Neto, H. T., and Almeida, V. (2019). Characterizing attention cascades in WhatsApp groups. In *Proceedings of the 11th International ACM Web Science Conference (WebSci’19)*, pages 27–36. Association for Computing Machinery (ACM).
- Calefato, F., Lanubile, F., Marasciulo, M. C., and Novielli, N. (2015). Mining successful answers in Stack Overflow. In *Proceedings of the 12th Working Conference on Mining Software Repositories*, pages 430–433. Institute of Electrical and Electronics Engineers (IEEE).

- Cambraia, C. N. (2013). Da lexicologia social a uma lexicologia sócio-histórica: Caminhos possíveis. *Revista de Estudos da Linguagem*, 21(1):157–188.
- Cangelosi, A. and Parisi, D., editors (2002). *Simulating the evolution of language*. Springer-Verlag, London.
- Carvalho, P., Sarmiento, L., Teixeira, J., and Silva, M. J. (2011). Liars and saviors in a sentiment annotated corpus of comments to political debates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 564–568. Association for Computational Linguistics (ACL).
- Collins Dictionary (2017). Etymology corner – Collins Word of the Year 2017. Retrieved from <https://www.collinsdictionary.com/word-lovers-blog/new/etymology-corner-collins-word-of-the-year-2017,400,HCB.html>. Accessed on December 22, 2018.
- Conroy, N. J., Rubin, V. L., and Chen, Y. (2016). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.
- Corrado, E. M. (2005). The importance of open access, open source, and open standards for libraries. *Issues in Science and Technology Librarianship*, 42.
- Correa, D., Silva, L. A., Mondal, M., Benevenuto, F., and Gumadi, K. P. (2015). The many shades of anonymity: Characterizing anonymous social media content. In *Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM-15)*, pages 71–80. Association for the Advancement of Artificial Intelligence (AAAI).

- Cotterell, R., Renduchintala, A., Saphra, N., and Callison-Burch, C. (2014). An Algerian Arabic-French code-switched corpus. In *Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools*, pages 34–37.
- Croft, W. (2000). *Explaining language change: An evolutionary approach*. Pearson Longman, Harlow.
- Crystal, D. (2004). *The language revolution*. Polity Press, Cambridge/Malden.
- Crystal, D. (2010). *The Cambridge encyclopedia of language*. Cambridge University Press, Cambridge, 3rd edition.
- Crystal, D. (2011). *Internet linguistics: A student guide*. Routledge, Abingdon.
- Culpeper, J. (2014). Keywords and characterization: An analysis of six characters in *Romeo and Juliet*. In Hoover, D. L., Culpeper, J., and O'Halloran, K., editors, *Digital literary studies: Corpus approaches to poetry, prose, and drama*, pages 9–34. Routledge.
- Cunha, E. L. T. P. (2012). Etiketagem de micromensagens no Twitter: Uma abordagem linguística. Master's thesis, Universidade Federal de Minas Gerais (UFMG).
- Cunha, E. L. T. P., Magno, G., and Almeida, V. (2017). A elaboração de um coletor e de um corpus de comentários extraídos de portais de notícias. In *Anais do X Congresso Internacional da ABRALIN*, pages 764–771. Associação Brasileira de Linguística (ABRALIN).
- Cunha, E. L. T. P., Magno, G., and Almeida, V. (under review). *Xereta: A Brazilian corpus of online news comments*. Manuscript submitted for publication.

- Cunha, E. L. T. P., Magno, G., Almeida, V., Gonçalves, M. A., and Benevenuto, F. (2012). A gender based study of tagging behavior in Twitter. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media (HT'12)*, pages 323–324. Association for Computing Machinery (ACM).
- Cunha, E. L. T. P., Magno, G., Caetano, J., Teixeira, D., and Almeida, V. (2018). Fake news as we feel it: Perception and conceptualization of the term “fake news” in the media. In Staab, S., Koltsova, O., and Ignatov, D., editors, *Social informatics*, volume 11185 of *Lecture Notes in Computer Science*, pages 151–166. Springer, Cham.
- Cunha, E. L. T. P., Magno, G., Caetano, J., Teixeira, D., and Almeida, V. (under review). *Quantifying the conceptualization of the term “fake news” in Brazilian and English-speaking media sources*. Manuscript submitted for publication.
- Cunha, E. L. T. P., Magno, G., Comarela, G., Almeida, V., Gonçalves, M. A., and Benevenuto, F. (2011). Analyzing the dynamic evolution of hashtags on Twitter: A language-based approach. In *Proceedings of the ACL Workshop on Language in Social Media (LSM 2011)*, pages 58–65. Association for Computational Linguistics (ACL).
- Cunha, E. L. T. P., Magno, G., Gonçalves, M. A., Cambraia, C., and Almeida, V. (2013). A linguistic characterization of Google+ posts across different social groups. In *Notes of the 5th Workshop on Information in Networks (WIN 2013)*. Stern School of Business – New York University (NYU).
- Cunha, E. L. T. P., Magno, G., Gonçalves, M. A., Cambraia, C., and Almeida, V. (2014a). He votes or she votes? Female and male discursive strategies in Twitter political hashtags. *PLOS ONE*, 9(1):e87041.

- Cunha, E. L. T. P., Magno, G., Gonçalves, M. A., Cambraia, C., and Almeida, V. (2014b). How you post is who you are: Characterizing Google+ status updates across social groups. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media (HT'14)*, pages 212–217. Association for Computing Machinery (ACM).
- Cunha, E. L. T. P. and Rocha, B. (2008). A escrita do jovem usuário de internet em contextos com motivação oral: Comparação com a formação histórica das línguas românicas e dos crioulos de base românica. In *Anais do VII Encontro de Lingüística de Corpus (ELC 2008)*. Universidade Estadual Paulista (UNESP).
- Cunha, E. L. T. P. and Wichmann, S. (in press). An algorithm to identify periods of establishment and obsolescence of linguistic items in a diachronic corpus. *Corpora*.
- Danescu-Niculescu-Mizil, C., Gamon, M., and Dumais, S. (2011). Mark my words! Linguistic style accommodation in social media. In *Proceedings of the 20th International World Wide Web Conference (WWW 2011)*, pages 745–754. Association for Computing Machinery (ACM).
- Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., and Potts, C. (2013). No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd International World Wide Web Conference (WWW 2013)*, pages 307–318. Association for Computing Machinery (ACM).
- Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, 7(2):121–157.
- Davies, M. (2013). Corpus of News on the Web (NOW): 3+ bil-

- lion words from 20 countries, updated every day. Retrieved from <https://corpus.byu.edu/now/> . Accessed on May 4, 2018.
- Davies, M. (2017a). Fake news. Retrieved from <https://corpus.byu.edu/now/help/fake-news.asp> . Accessed on May 4, 2018.
- Davies, M. (2017b). The new 4.3 billion word NOW corpus, with 4-5 million words of data added every day. In *Proceedings of the 9th International Corpus Linguistics Conference*.
- Day, M. (2006). The long-term preservation of Web content. In Masanès, J., editor, *Web archiving*, pages 177–199. Springer, Berlin.
- De Choudhury, M., Sundaram, H., John, A., and Seligmann, D. D. (2009). What makes conversations interesting? Themes, participants and consequences of conversations in online social media. In *Proceedings of the 18th International Conference on World Wide Web (WWW 2009)*, pages 331–340. Association for Computing Machinery (ACM).
- Degand, L. and Van Bergen, G. (2018). Discourse markers as turn-transition devices: Evidence from speech and instant messaging. *Discourse Processes*, 55(1):47–71.
- Dores, M. and Toledo, C. (2018). De “lepra” à “hanseníase”: Uma análise lexicológica de base sócio-histórica. *Diacrítica*, 32(1):179–208.
- Dubois, J., Giacomo, M., Guespin, L., Marcellesi, C., and Mevel, J.-P. (1986). *Dicionário de linguística*. Cultrix, São Paulo, 2nd edition.
- Dunst, A., Hartel, R., and Laubrock, J. (2017). The Graphic Narrative Corpus (GNC): Design, annotation, and analysis for the

digital humanities. In *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pages 15–20. Institute of Electrical and Electronics Engineers (IEEE).

Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2014). Diffusion of lexical change in social media. *PLOS ONE*, 9(11):e113114.

Erjavec, K. and Kovačič, M. P. (2012). “You don’t understand, this is a new war!” Analysis of hate speech in news web sites’ comments. *Mass Communication and Society*, 15(6):899–920.

Fast, E., Chen, B., and Bernstein, M. S. (2016). Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657. Association for Computing Machinery (ACM).

Ferguson, C., Inglis, S. C., Newton, P. J., Cripps, P. J. S., Macdonald, P. S., and Davidson, P. M. (2014). Social media: A tool to spread information: A case study analysis of Twitter conversation at the Cardiac Society of Australia & New Zealand 61st Annual Scientific Meeting 2013. *Collegian*, 21(2):89–93.

Fišer, D., Ljubešić, N., and Erjavec, T. (2018). The Janes project: Language resources and tools for Slovene user generated content. *Language Resources & Evaluation*.

Flaounas, I., Ali, O., Lansdall-Welfare, T., Bie, T. D., Mosdell, N., Lewis, J., and Cristianini, N. (2013). Research methods in the age of digital journalism: Massive-scale automated analysis of news-content – topics, style and gender. *Digital Journalism*, 1(1):102–116.

- Flaounas, I., Turchi, M., Ali, O., Fyson, N., De Bie, T., Mosdell, N., Lewis, J., and Cristianini, N. (2010). The structure of the EU mediasphere. *PLOS ONE*, 5(12):e14243.
- Freire, A., Lloyd, R., and Turgeon, M. (2017). “Seu petralha! Seu coxinha!” – Measuring affective polarization in Brazil. In *Proceedings of the 41o. Encontro Anual da ANPOCS*. Associação Nacional dos Pesquisadores de Ciências Sociais (ANPOCS).
- Frischer, B. (2011). Art and science in the age of digital reproduction: From mimetic representation to interactive virtual reality. *Virtual Archaeology Review*, 2(4):19–32.
- García-Calderón, M. Á., García-Hernández, R. A., and Ledeneva, Y. (2018). Unsupervised multi-language handwritten text line segmentation. *Journal of Intelligent & Fuzzy Systems*, 34(5):2901–2911.
- Gelfert, A. (2018). Fake news: A definition. *Informal Logic*, 38(1):84–117.
- Glaznieks, A., Nicolas, L., Stemle, E., Abel, A., and Lyding, V. (2014). Establishing a standardised procedure for building learner corpora. *Apples – Journal of Applied Language Studies*, 8(3):5–20.
- Glowacki, M., Narayanan, V., Maynard, S., Hirsch, G., Kollanyi, B., Neudert, L.-M., Howard, P., Lederer, T., and Barash, V. (2018). News and political information consumption in Mexico: Mapping the 2018 Mexican presidential election on Twitter and Facebook. *The Computational Propaganda Project: Algorithms, automation and digital politics (COMPROP)*.
- Gómez, V., Kaltenbrunner, A., and López, V. (2008). Statistical analysis of the social network and discussion threads in Slashdot.

In *Proceedings of the 17th International Conference on World Wide Web (WWW 2008)*, pages 645–654. Association for Computing Machinery (ACM).

Gries, S. T. and Hilpert, M. (2008). The identification of stages in diachronic data: Variability-based neighbour clustering. *Corpora*, 3(1):59–81.

Gross, M. (1972). *Mathematical models in linguistics*. Prentice-Hall, Englewood Cliffs.

Grosseck, G. (2009). To use or not to use web 2.0 in higher education? *Procedia Social and Behavioral Sciences*, 1:478–482.

Grzybek, P. (2007). History and methodology of word length studies. In Grzybek, P., editor, *Contributions to the science of text and language: Word length studies and related issues*, pages 15–90. Springer, Berlin.

Guedes, A. d. S. and Mendes, B. P. (2016). Um estudo lexicológico de base sócio-histórica das formas lexicais “asilo de idosos” e “casa de repouso”. *Raído*, 10(24):38–52.

Gulordava, K. and Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, pages 67–71. Association for Computational Linguistics (ACL).

Habgood-Coote, J. (2018). The term ‘fake news’ is doing great harm. The Conversation. Retrieved from <http://theconversation.com/the-term-fake-news-is-doing-great-harm-100406> . Accessed on January 5, 2019.

- Habgood-Coote, J. (2019). Stop talking about fake news! *Inquiry*, 62(9-10):1033–1065.
- Halabi, Y. S., Sa, Z., Hamdan, F., and Yousef, K. H. (2009). Modeling adaptive degraded document image binarization and optical character system. *European Journal of Scientific Research*, 28(1):14–32.
- Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1501. Association for Computational Linguistics (ACL).
- Harden, C. (2019). Brazil fell for fake news: What to do about it now? Retrieved from <https://www.wilsoncenter.org/blog-post/brazil-fell-for-fake-news-what-to-do-about-it-now>. Accessed on August 12, 2019.
- Heaps, H. S. (1978). *Information retrieval: Computational and theoretical aspects*. Academic Press, New York.
- Heeringa, W. and Prokić, J. (2017). Computational dialectology. In Boberg, C., Nerbonne, J., and Watt, D., editors, *The handbook of dialectology*, pages 330–347. John Wiley & Sons, Hoboken.
- Henrich, N. and Holmes, B. (2013). Web news readers’ comments: Towards developing a methodology for using on-line comments in social inquiry. *Journal of Media and Communication Studies*, 5(1):1–4.
- Herdan, G. (1964). *Quantitative linguistics*. Butterworth, London.
- Hilpert, M. and Gries, S. T. (2009). Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus

linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24(4):385–401.

Hilpert, M. and Gries, S. T. (2016). Quantitative approaches to diachronic corpus linguistics. In Kytö, M. and Pahta, P., editors, *The Cambridge handbook of English historical linguistics*, pages 36–53. Cambridge University Press, Cambridge.

Hilpert, M. and Mair, C. (2015). Grammatical change. In Biber, D. and Reppen, R., editors, *The Cambridge handbook of English corpus linguistics*, pages 180–200. Cambridge University Press, Cambridge.

Hinrichs, L. and Szmrecsanyi, B. (2007). Recent changes in the function and frequency of Standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language and Linguistics*, 11(3):437–474.

Holan, A. (2017). The media’s definition of fake news vs. Donald Trump’s. *First Amendment Law Review*, 16:121–128.

Hooker, C. (2019). Students and corporate social media: Do college students care about social media usernames? *International Journal of Social Science Studies*, 7(4):79–86.

Horta Ribeiro, M., Calais, P. H., Almeida, V. A., and Meira Jr., W. (2017). “Everything I disagree with is #FakeNews”: Correlating political polarization and spread of misinformation. In *Proceedings of Data Science + Journalism @ KDD 2017 (DS+J’17)*.

Hsu, C.-F., Khabiri, E., and Caverlee, J. (2009). Ranking comments on the social web. In *2009 International Conference on Computational Science and Engineering*, volume 4, pages 90–97. Institute of Electrical and Electronics Engineers (IEEE).

- Hundt, M. and Mair, C. (1999). “Agile” and “uptight” genres: The corpus-based approach to language change in progress. *International Journal of Corpus Linguistics*, 4:221–242.
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press, Cambridge.
- Hunter, W. and Power, T. J. (2019). Bolsonaro and Brazil’s illiberal backlash. *Journal of Democracy*, 30(1):68–82.
- Jensen, K. E. (2014). Linguistics in the digital humanities: (Computational) corpus linguistics. *MedieKultur: Journal of Media and Communication Research*, 57:115–134.
- Kerremans, D., Stegmayr, S., and Schmid, H.-J. (2012). The NeoCrawler: Identifying and retrieving neologisms from the Internet and monitoring ongoing change. In Allan, K. and Robinson, J. A., editors, *Current methods in historical semantics*, pages 59–96. De Gruyter Mouton, Berlin.
- Kilgariff, A. (2005). Web as corpus (2001). In Sampson, G. and McCarthy, D., editors, *Corpus linguistics: Readings in a widening discipline*, pages 471–473. Continuum, London.
- Knowles, E. and Elliott, J., editors (1997). *The Oxford dictionary of new words*. Oxford University Press, Oxford.
- Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., and Taboada, M. (2019). The SFU Opinion and Comments Corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*.
- Koplenig, A. (2017). The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets – Reconstructing the composition of the German

corpus in times of WWII. *Digital Scholarship in the Humanities*, 32(1):169–188.

Krishnamurthy, R. (1997). Keeping good company: Collocation, corpus and dictionaries. In *Cicle de conferències 95-96: Lèxic, corpus i diccionaris*, Barcelona. Institut Universitari de Lingüística Aplicada/Universitat Pompeu Fabra.

Krishnamurthy, R. (2003). Freeze-frame pictures: Micro-diachronic variations in synchronic corpora. In Jozsef Andor, J. H. and Nikolov, M., editors, *Studies in English theoretical and applied linguistics*, pages 15–31. Lingua Franca Csoport, Pecs.

Küng, L., Newman, N., and Picard, R. G. (2016). Online news. In Bauer, J. M. and Latzer, M., editors, *Handbook on the economics of the Internet*, pages 443–457. Edward Elgar Publishing, Cheltenham/Northampton.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites*, volume 1. Stanford University Press.

Lansdall-Welfare, T., Sudhahar, S., Veltri, G. A., and Cristianini, N. (2014). On the coverage of science in the media: A big data study on the impact of the Fukushima disaster. In *2014 IEEE International Conference on Big Data*, pages 60–66.

Las Casas, D., Magno, G., Cunha, E. L. T. P., Gonçalves, M. A., Cambraia, C., and Almeida, V. (2014). Noticing the other gender on Google+. In *Proceedings of the ACM Web Science 2014 Conference (WebSci 2014)*, pages 156–160. Association for Computing Machinery (ACM).

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R.,

- Thorson, E. A., Watts, D. J., and Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380):1094–1096.
- Lee, E.-J. (2012). That’s not the way it is: How user-generated comments on the news affect perceived media bias. *Journal of Computer-Mediated Communication*, 18:32–45.
- Leech, G. (2005). Adding linguistic annotation. In Wynne, M., editor, *Developing linguistic corpora: A guide to good practice*, pages 17–29. Oxbow Books, Oxford. Retrieved from <http://ota.ox.ac.uk/documents/creating/dlc/>. Accessed on August 10, 2019.
- Lees, C. (2018). Fake news: The global silencer. *Index on Censorship*, 47(1):88–91.
- Leetaru, K. (2011). Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, 16(9).
- Lüdeling, A. and Kytö, M., editors (2008). *Corpus linguistics: An international handbook*, volume 1. De Gruyter, Berlin/New York.
- Maia, R. C. M., Rossini, P. G. C., de Oliveira, V. V., and de Oliveira, A. G. (2015). Sobre a importância de examinar diferentes ambientes online em estudos de deliberação. *Opinião Pública*, 21(2):490–513.
- Matoré, G. (1949). La lexicologie sociale. *L’Information Littéraire*, 2.
- Matoré, G. (1953). *La méthode en lexicologie: Domaine français*. Didier, Paris.
- McEnery, T. and Wilson, A. (1996). *Corpus linguistics: An introduction*. Edinburgh University Press, Edinburgh.

- Mellish, C. (1994). Computers and language use. In Asher, R. and Simpson, J., editors, *The encyclopedia of language and linguistics*, volume 2, pages 672–673. Pergamon Press, Oxford.
- Mello, H. (2012). Os corpora orais e o C-ORAL-BRASIL. In Raso, T. and Mello, H., editors, *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*, pages 31–54. Editora UFMG, Belo Horizonte.
- Mello, H., Raso, T., Mittmann, M. M., Vale, H. P., and Côrtes, P. O. (2012). Transcrição e segmentação prosódica do corpus C-ORAL-BRASIL: Critérios de implementação e validação. In Raso, T. and Mello, H., editors, *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*, pages 125–176. Editora UFMG, Belo Horizonte.
- Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., and Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.
- Milioni, D. L., Vadratsikas, K., and Papa, V. (2012). ‘Their two cents worth’: Exploring user agency in readers’ comments in on-line news media. *Observatorio (OBS*) Journal*, 6(3):21–47.
- Mitkov, R., editor (2005). *The Oxford handbook of computational linguistics*. Oxford University Press.
- Möller, A. M., Kühne, R., Baumgartner, S. E., and Peter, J. (2019). Exploring user responses to entertainment and political videos: An automated content analysis of YouTube. *Social Science Computer Review*, 37(4):510–528.

- Moon, R. (2010). What can a corpus tell us about lexis? In O’Keeffe, A. and McCarthy, M., editors, *The Routledge handbook of corpus linguistics*, pages 197–211. Routledge, New York.
- Moreira, T., Brigatto, G., and Rosa, J. L. (2018). Após PagSeguro, UOL busca investidores para outros negócios. Retrieved from <https://glo.bo/2FD1MT1> . Accessed on August 15, 2019.
- Napoles, C., Tetreault, J., Rosato, E., Provenzale, B., and Pappu, A. (2017). Finding good conversations online: The Yahoo News Annotated Comments Corpus. In *Proceedings of the 11th Linguistic Annotation Workshop (LAW-XI)*, pages 13–23. Association for Computational Linguistics (ACL).
- Nettle, D. (1999). Using social impact theory to simulate language change. *Lingua*, 108:95–117.
- Nevalainen, T. and Raumolin-Brunberg, H., editors (1996). *Sociolinguistics and language history: Studies based on the Corpus of Early English Correspondence*. Rodopi, Amsterdam/Atlanta.
- Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, pages 100–108. Association for Computational Linguistics (ACL).
- Nguyen, D., Doğruöz, A. S., Rosé, C. P., and de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational Linguistics*, 42(3):537–593.
- Nielsen, R. K. and Graves, L. (2017). “News you don’t believe”: Audience perspectives on fake news. Retrieved from <https://reutersinstitute.politics.ox.ac.uk/>

sites/default/files/2017-10/Nielsen&Graves_factsheet_1710v3_FINAL_download.pdf . Accessed on July 02, 2019.

O'Reilly, T. (2007). What is Web 2.0: Design patterns and business models for the next generation of software. *Communications & Strategies*, 65:17–37.

Ottoni, R., Cunha, E. L. T. P., Magno, G., Bernardina, P., Meira Jr., W., and Almeida, V. (2018). Analyzing right-wing YouTube channels: Hate, violence and discrimination. In *Proceedings of the 10th ACM Conference on Web Science (WebSci'18)*, pages 323–332. Association for Computing Machinery (ACM).

Pagel, M., Atkinson, Q. D., and Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, 449:717–720.

Parliament of the United Kingdom (2018). Disinformation and ‘fake news’: Interim report. Retrieved from <https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/363/36304.htm> . Accessed on January 2, 2018.

Partington, A. (2010). Modern diachronic corpus-assisted discourse studies (MD-CADS) on UK newspapers: An overview of the project. *Corpora*, 5(2):83–108.

Pate, U. A. and Ibrahim, A. M. (2019). Fake news, hate speech and Nigeria’s struggle for democratic consolidation: A conceptual review. In Solo, A. M., editor, *Handbook of research on politics in the computer age*, pages 89–112. IGI Global, Hershey.

Pechenick, E. A., Danforth, C. M., and Dodds, P. S. (2015). Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLOS ONE*, 10(10):e0137041.

- Pedro, G. W. (2018). ComentCorpus: Identificação e pistas linguísticas para detecção de ironia no português do Brasil. Master's thesis, Universidade Federal de São Carlos (UFSCar).
- Pennebaker, J. W., Booth, R. J., Boyd, R. L., and Francis, M. E. (2015). *Linguistic Inquiry and Word Count: LIWC2015*. Pennebaker Conglomerates, Austin.
- Perc, M. (2012). Evolution of the most common English words and phrases over the centuries. *Journal of The Royal Society Interface*, 9:3323–3328.
- Petersen, A. M., Tenenbaum, J., Havlin, S., and Stanley, H. E. (2012). Statistical laws governing fluctuations in word use from word birth to word death. *Scientific Reports*, 2(313).
- Phillips, T. (2018). Bolsonaro business backers accused of illegal Whatsapp fake news campaign. The Guardian. Retrieved from <https://www.theguardian.com/world/2018/oct/18/brazil-jair-bolsonaro-whatsapp-fake-news-campaign>. Accessed on January 2, 2018.
- Potthast, M. (2009). Measuring the descriptiveness of Web comments. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09)*, pages 724–725. Association for Computing Machinery (ACM).
- Potthast, M. and Becker, S. (2010). Opinion summarization of Web comments. In Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., and van Rijsbergen, K., editors, *Advances in information retrieval: Proceedings of the 32nd European Conference on Information Retrieval (ECIR 2010)*, pages 668–669. Springer.

- Presner, T. and Johanson, C. (2009). The promise of Digital Humanities: A whitepaper. Retrieved from <http://humanitiesblast.com/Promise%20of%20Digital%20Humanities.pdf> . Accessed on July 02, 2019.
- Prokić, J. (2017). Quantitative diachronic dialectology. In Wieling, M., Kroon, M., van Noord, G., and Bouma, G., editors, *From semantics to dialectometry: Festschrift in honour of John Nerbonne*, pages 293–301. College Publications, London.
- Rafael, G. C. R. A. and Simião, D. P. (2019). Aidético e soropositivo: Análise sócio-histórica da concorrência entre qualificadores utilizados em referência a portadores do HIV. *Inventário*, 23:45–68.
- Rajaraman, A. and Ullman, J. D. (2011). *Mining of massive datasets*. Cambridge University Press.
- Rama, T. (2015). *Studies in computational historical linguistics: Models and analyses*. PhD thesis, University of Gothenburg.
- Ramalho, R. (2018). Fux diz que Justiça pode anular uma eleição se resultado for influenciado por ‘fake news’ em massa. Retrieved from <https://glo.bo/36JPmVh> . Accessed on June 28, 2019.
- Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., and Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMMLP)*, pages 2931–2937. Association for Computational Linguistics (ACL).
- Řehůřek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. ELRA.

- Reis, J. C., Correia, A., Murai, F., Veloso, A., and Benevenuto, F. (2019). Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81.
- Reyes, A., Potthast, M., Rosso, P., and Stein, B. (2010). Evaluating humour features on Web comments. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC’10)*. European Language Resources Association (ELRA).
- Ross, A. S. and Rivers, D. J. (2018). Discursive deflection: Accusation of “fake news” and the spread of mis- and disinformation in the tweets of President Trump. *Social Media + Society*, 4(2):1–12.
- Rossini, P. G. C. (2017). *Conversação política, incivilidade e intolerância em ambientes digitais*. PhD thesis, Universidade Federal de Minas Gerais.
- Roth, S. (2014). Fashionable functions: A Google Ngram view of trends in functional differentiation (1800-2000). *International Journal of Technology and Human Interaction*, 10(2):34–58.
- Rubin, V. L., Conroy, N. J., Chen, Y., and Cornwell, S. (2016). Fake news or truth? Using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17. Association for Computational Linguistics (ACL).
- Ruchansky, N., Seo, S., and Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM Conference on Information and Knowledge Management (CIKM’17)*, pages 797–806. Association for Computing Machinery (ACM).

- Rumšienė, G. (2004). Development of Internet English: Alternative lexis, syntax and morphology. *Kalby Studijos (Studies about Languages)*, 6:48–55.
- Saccenti, E. and Tenori, L. (2012). Stylometric investigation of Dante’s *Divina Commedia* by means of multivariate data analysis techniques. *International Journal of Computational Linguistics Research*, 3(2):35–48.
- Saliés, T. G. and Shepherd, T. G. (2013). Introdução: Por uma Linguística da Internet. In Shepherd, T. G. and Saliés, T. G., editors, *Linguística da internet*, pages 7–14. Contexto, São Paulo.
- Sampson, G. and McCarthy, D. (2005). From *The structure of English* (1952). In Sampson, G. and McCarthy, D., editors, *Corpus linguistics: Readings in a widening discipline*, page 9. Continuum, London.
- Sarmiento, R. and Mendonça, R. F. (2016). Disrespect in online deliberation: Inducing factors and democratic potentials. *Revista de Ciencia Política (Santiago)*, 36(3):705–729.
- Saussure, F. (1916). *Cours de linguistique générale*. Published by Charles Bally and Albert Sechehaye. Payot, Lausanne/Paris.
- Schmid, H.-J. (2007). Entrenchment, salience, and basic levels. In Geeraerts, D. and Cuyckens, H., editors, *The Oxford handbook of cognitive linguistics*, pages 117–138. Oxford University Press, Oxford.
- Schulze, C., Stauffer, D., and Wichmann, S. (2008). Birth, survival and death of languages by Monte Carlo simulation. *Communications in Computational Physics*, 3(2):271–294.

- Sebba, M. and Fligelstone, S. (1994). Corpora. In Asher, R. and Simpson, J., editors, *The encyclopedia of language and linguistics*, volume 2, pages 769–773. Pergamon Press, Oxford.
- Shah, C. and Pomerantz, J. (2010). Evaluating and predicting answer quality in community QA. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 411–418. Association for Computing Machinery (ACM).
- Shepherd, T. M. G. (2014). Changing ‘faces’: A case study of complex prepositions in Brazilian Portuguese. In Berber Sardinha, T. and São Bento Ferreira, T. L., editors, *Working with Portuguese Corpora*, pages 69–88. Bloomsbury Academic.
- Shertzer, M. D. (1996). *The elements of grammar*. Macmillan Publishing Company, New York.
- Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.
- Siersdorfer, S., Chelaru, S., Nejdl, W., and San Pedro, J. (2010). How useful are your comments? Analyzing and predicting YouTube comments and comment ratings. In *Proceedings of the 19th International Conference on World Wide Web (WWW 2010)*, pages 891–900. Association for Computing Machinery (ACM).
- Silva, N. F. F., Coletta, L. F. S., and Hruschka, E. R. (2016). A survey and comparative study of tweet sentiment analysis via semi-supervised learning. *ACM Computing Surveys*, 49(1).
- SimilarWeb (2016). Rank. Retrieved from <https://support.similarweb.com/hc/en-us/articles/213452305-Rank->. Accessed on August 26, 2019.

- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press, Oxford.
- Standage, T. (2017). The true history of fake news. 1843 Magazine. Retrieved from <https://www.1843magazine.com/technology/rewind/the-true-history-of-fake-news> . Accessed on May 4, 2018.
- Sullivan, M. (2017). It's time to retire the tainted term 'fake news'. The Washington Post. Retrieved from https://www.washingtonpost.com/lifestyle/style/its-time-to-retire-the-tainted-term-fake-news/2017/01/06/a5a7516c-d375-11e6-945a-76f69a399dd5_story.html . Accessed on January 5, 2019.
- Svensén, B. (1993). *Practical lexicography: Principles and methods of dictionary-making*. Oxford University Press, Oxford.
- Tambini, D. (2017). Fake news: Public policy responses. In Tambini, D. and Goodman, E., editors, *LSE Media Policy Project Series*. Media Policy Project, London.
- Tandoc Jr., E. C., Lim, Z. W., and Ling, R. (2018). Defining “fake news”: A typology of scholarly definitions. *Digital Journalism*, 6(2):137–153.
- Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.
- Taylor, C. (2008). What is *corpus linguistics*? What the data says. *ICAME Journal*, 32:179–200.
- Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. (2010). Sentiment in short strength detection informal text. *Jour-*

nal of the American Society for Information Science and Technology, 61(12):2544–2558.

Thompson, S. E. and Parthasarathy, S. (2006). Moore’s law: The future of Si microelectronics. *Materials Today*, 9(6):20–25.

Tichý, O. (2018). Lexical obsolescence and loss in English: 1700–2000. In Kopaczyk, J. and Tyrkkö, J., editors, *Applications of pattern-driven methods in corpus linguistics*, pages 81–103. John Benjamins, Amsterdam.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003*. Association for Computational Linguistics (ACL).

Trevisani, M. and Tuzzi, A. (2018). Learning the evolution of disciplines from scientific literature: A functional clustering approach to normalized keyword count trajectories. *Knowledge-Based Systems*, 146:129–141.

Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., and Krause, A. (2018). Fake news detection in social networks via crowd signals. In *Proceedings of the Web Conference (TheWebConf 2018)*, pages 517–524. Association for Computing Machinery (ACM).

Tulloch, S. (1991). *The Oxford dictionary of new words: A popular guide to words in the news*. Oxford University Press, Oxford.

Upadhyay, S., Pant, V., Bhasin, S., and Pattanshetti, M. K. (2017). Articulating the construction of a web scraper for massive data extraction. In *Proceedings of the 2017 Second International Conference on Electrical, Computer and Communication Technolo-*

- gies (ICECCT)*. Institute of Electrical and Electronics Engineers (IEEE).
- Van Hout, T. and Burger, P. (2015). Mediatization and the language of journalism. *Tilburg Papers in Culture Studies*, 131.
- Vicario, M. D., Quattrociocchi, W., Scala, A., and Zollo, F. (2019). Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web (TWEB)*, 13(2).
- Vinodhini, G. and Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: A survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(6):282–292.
- Vo, A. T. and Carter, R. (2010). What can a corpus tell us about creativity? In O’Keeffe, A. and McCarthy, M., editors, *The Routledge handbook of corpus linguistics*, pages 302–315. Routledge, New York.
- Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.
- Wan, X., Zong, L., Huang, X., Ma, T., Jia, H., Wu, Y., and Xiao, J. (2011). Named entity recognition in Chinese news comments on the Web. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 856–864.
- Wasserman, H. (2017). Fake news from Africa: Panics, politics and paradigm. *Journalism*.
- Webster, P. (2017). Users, technologies, organisations: Towards a cultural history of world web archiving. In Brügger, N., editor, *Web 25: Histories from 25 years of the World Wide Web*, pages 179–190. Peter Lang, New York.

- Widdowson, H. G. (2000). On the limitations of linguistics applied. *Applied Linguistics*, 21(1):3–25.
- Wintner, S. (2010). Computational models of language acquisition. In Gelbukh, A., editor, *Computational linguistics and intelligent text processing*, volume 6008 of *Lecture Notes in Computer Science*, pages 86–99. Springer, Berlin and Heidelberg.
- Woolfs, D. (2010). Computational forensic linguistics: Searching for similarity in large specialised corpora. In Coulthard, M. and Johnson, A., editors, *The Routledge handbook of forensic linguistics*, pages 576–590. Routledge, New York.
- Wynne, M. (2008). Searching and concordancing. In Lüdeling, A. and Kytö, M., editors, *Corpus linguistics: An international handbook*, volume 1, pages 706–737. De Gruyter, Berlin/New York.
- Zannettou, S., Caulfield, T., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Sirivianos, M., Stringhini, G., and Blackburn, J. (2017). The Web centipede: Understanding how Web communities influence each other through the lens of mainstream and alternative news sources. In *Proceedings of the 2017 Internet Measurement Conference (IMC 2017)*, pages 405–417. Association for Computing Machinery (ACM).
- Zhang, W., Wu, F., and Zhang, C. (2013). Interpretation of the formation of internet neologisms and their translation from Pound’s perspective of “language energy”. *International Journal of English Linguistics*, 3(2):66–71.
- Zinsmeister, H. (2015). Syntax and corpora. In Kiss, T. and Alexiadou, A., editors, *Syntax – theory and analysis: An international handbook*, volume 3, pages 1912–1941. De Gruyter Mouton, Berlin/Munich/Boston.

Zollo, F., Novak, P. K., Del Vicario, M., Bessi, A., Mozetič, I., Scala, A., Caldarelli, G., and Quattrociocchi, W. (2015). Emotional dynamics in the age of misinformation. *PLOS ONE*, 10(9).

