Vol. 149, No. 1 · Short communication

DIE ERDE

Journal of the
Geographical Society
of Berlin

Quantitative text analysis in Geography: facilitating access and fostering collaboration

Thilo Wiertz

Institut für Umweltsozialwissenschaften und Geographie, Albert-Ludwigs-Universität Freiburg, 79085 Freiburg, thilo.wiertz@geographie.uni-freiburg.de

Manuscript submitted: 19 December 2017 / Accepted for publication: 15 January 2018 / Published online: 30 March 2018

Abstract

Quantitative text analysis can support researchers working with a large number of documents. Corpus linguistic methods are already employed by geographers in the context of discourse studies, and recent discussions about big data and digital geographies point to a potential broadening of their application. However, building a corpus and adapting to existing data analysis tools can be challenging. In this paper, we outline possible steps towards collaborative quantitative text analysis through the use of computational methods and corpora that can be incorporated into a variety of research settings. We summarise key steps for creating annotated corpora from text sources using state of the art methods and tools. Using the open source software Corpus Workbench (Evert and Hardie 2011) and CQPweb (Hardie 2012), we provide a platform to access corpora and corpus analysis functionality via a web interface. We invite researchers to use this existing infrastructure for corpus linguistic methods in their teaching and research, and to collaborate in making interesting material available to the geographic research community.

Zusammenfassung

Quantitative Methoden können ForscherInnen bei der Auswertung umfangreicher Textquellen unterstützen. Korpuslinguistische Verfahren werden im Bereich der humangeographischen Diskursforschung bereits seit einiger Zeit eingesetzt, und aktuelle Diskussionen um Big Data und Digitale Geographien verweisen auf neue denkbare Anwendungsfelder. Einer breiteren Anwendung entsprechender Methoden in der Forschung stehen jedoch häufig der Aufwand der Korpuserstellung sowie die Einarbeitung in Analysesoftware entgegen. Der vorliegende Beitrag stellt neuere Techniken der kollaborativen Arbeit mit Textkorpora vor, die es ermöglichen, quantitative Verfahren in eine Vielzahl von Forschungskontexten einzubeziehen. Wir beschreiben, wie sich Textquellen mithilfe aktueller Methoden und Werkzeuge in annotierte Korpora überführen lassen. Unter Verwendung der OpenSource-Software Corpus Workbench (*Evert* and *Hardie* 2011) sowie CQPweb (*Hardie* 2012) stellen wir eine Online-Plattform bereit, die es erlaubt, über eine Weboberfläche auf Korpora und korpuslinguistische Analysemethoden zuzugreifen. Wir laden interessierte Forscherinnen und Forscher ein, die bestehende Plattform im Rahmen von Forschung und Lehre zu nutzen und an der Erschließung und Bereitstellung gemeinschaftlich genutzter Korpora mitzuwirken.

Keywords human geography, digital geography, quantitative methods, corpus linguistics, discourse analysis, big data

Thilo Wiertz 2018: Quantitative text analysis in Geography: facilitating access and fostering collaboration. – DIE ERDE **149** (1): 52-56

DOI: 10.12854/erde-149-58

1. Corpus linguistic methods in Geography

Methods for quantitative text analysis can support research wherever large numbers of documents hold potentially interesting information about a topic. In the context of social science, 'large' typically means at least hundreds or thousands of documents, an amount that cannot be surveyed by a research team using qualitative methods alone. Methods developed in corpus linguistics can support the analysis of such collections by identifying statistically significant patterns within the material (*Bubenhofer* 2009: 16). Simple applications with geographic relevance could be, for instance, identifying terms that co-occur frequently with a concept of interest, such as 'Frankfurt' or 'climate', or words that are characteristic to news reports about a particular country. Similarly, temporal shifts in political or public discussions about topics such as migration or sustainable development could be identified. In Geography, such methods have primarily been employed in the context of discourse analysis (*Dzudzek* et al. 2009; *Glasze* 2008; *Mattissek* and *Glasze* 2016), contributing to three main research themes: (i) Analyses of (geo)political images ask how geographic entities, such as a region or a state, are ascribed particular attributes and characteristics that in turn shape socio-political relations and actions (*Linnemann* and *Reuber* 2015); (ii) the identity of subjects can be understood and analysed as the result of discursive distinctions between Self and Other, distinctions that are often made in reference to geography (*Glasze* 2013; *Glasze* and *Weber* 2014); (iii) processes of geographic interest, for instance urban governance and development, follow particular ways of framing problems and justifying actions. The analysis of such political rationalities is a key concern for geographic studies of governmentality (*Mattissek* 2008; *Mattissek* and *Sturm* 2017). These research studies employ corpus linguistic methods as a first step to identify regularities of association that become visible across a large section of the material, and then proceed with a qualitative analysis of selected texts or passages.

Despite existing experience with corpus linguistic methods in Geography, practical and technical challenges still pose a barrier to more extensive adoption. Preparing a corpus can be tedious and time consuming, and if only used within one specific project, the effort may appear prohibitive. Furthermore, while a variety of software tools for quantitative text analysis are available today, not all are well maintained, documented and accessible to non-experts. On the other hand, the continuous digitisation of socio-political communication, as well as a growing interest amongst geographers in big data and the digital sphere, provide good reasons to further explore the potentials of quantitative text analysis (cf. *Ash* et al. 2018; *DeLyser* and *Sui* 2013; *Graham* and *Shelton* 2013). We therefore suggest further collaborative efforts in this area. Computational methods and tools facilitate the creation of comprehensive corpora that can be used beyond an individual project. We briefly outline the key steps and standards relevant for creating a corpus from text sources. To make corpora and analysis functions accessible to a wider community of researchers, we set up a server running the open source software Corpus Workbench (*Evert* and *Hardie* 2011) and CQPweb (*Hardie* 2012). This makes corpora and corpus linguistic functionality accessible through a web browser and allows for the performance of more complex analytic tasks through a command line and 'R' interface.

2. Transforming text into collaborative research corpora

A variety of sources are of potential interest to geographic inquiries, such as, for instance, official and unofficial reports and documents, public or parliamentary debates, blogs, interviews or news reports. The increasing availability of texts and documents in structured digital formats facilitates the creation of comprehensive and collaborative corpora: Whereas corpus linguistic research in the social sciences often works with relatively small thematic corpora, comprehensiveness means that the collected and processed material is not unnecessarily restricted, e.g., to texts containing a particular keyword or a short time frame. Comprehensive corpora can be useful to a variety of researchers, and can help them answer a wide range of research questions. An exemplary effort in this respect is made by the Polmine project, which recently released a corpus of German parliamentary debates under a public license (*Blaette* 2017a). If comprehensive corpora become more widely available, corpus linguistic endeavours will not always have to start by building a new corpus. Where corpora are not (yet) available, a variety of computational tools and resources greatly facilitate the collection and extraction of material from documents, the processing and annotating of text, and the conversion into different (standardised) formats for analysis and exchange.

The following paragraphs provide a brief technical overview of these steps.

The first step in creating a corpus is extracting text from existing documents. Analogue documents need to be digitised first, and although this will almost always require manual interventions, computational tools can facilitate this task. Furthermore, a lot of potentially interesting material is available online and comes in structured digital file formats, such as HTML or XML. Large amounts of such material can be 'scraped' and processed relatively easily using programming environments such as 'R' or Python. *Munzert* et al. (2015) offer a helpful and very accessible introduction to web scraping and give examples for applications in the social sciences. Once a strategy for collecting a source and extracting the text is in place, the text usually needs to be cleaned, i.e., freed from irrelevant information, errors and artefacts, and tokenised, i.e., split into units of analysis, usually words and punctuation. Free text editors with support for regular expressions are capable of performing quite complex cleaning operations and can handle large amounts of text much more efficiently than standard office software. More complex tasks and materials are handled more easily in programming environments. While this may initially seem like a hurdle, the benefits, like reusing tools for future research and automating repetitive and time consuming manual data processing tasks, can pay off very quickly. 'R' has become very popular for text processing tasks in recent years, and offers a range of packages tailored to text manipulation and analysis (*Gries* 2009; *Jockers* 2014; *Silge* 2017). Given its statistic functionality and GIS extensions, it is also increasingly used in Geography research and teaching.

Computational linguistics has developed methods and tools for text annotations that can be very valuable, also for social scientific and geographic analyses. Part-of-speech tags complement each token with information about its grammatical role. They can be used, for example, to identify attributive adjectives that co-occur with a place name or another concept of interest. Lemmatisation annotates all flections of a word (e.g. migrates/migrated/migrating) with the corresponding base form or lemma (migrate), considerably enhancing the quality of search results and statistical analysis. Advances are also made in the automated identification of named entities, such as place names, persons or organisations, which may be particularly valuable to social science and geographic inquiries. Various command line tools are available for tagging and annotating text, notably the TreeTagger (*Schmid* 1999) that provides tokenisation, lemmatisation, and part of speech tagging for several languages, and the Stanford CoreNLP framework (*Manning* et al. 2014) that offers named entity recognition and identification of further grammatical relationships.

When processing documents, a decision has to be made on how to store the corpus. Often, the choice of data structure has followed the particular requirements of the software tool used for the analysis. This has led to a proliferation of various incompatible formats, hindering the exchange of corpora between different research environments. Consequently, standardisation has become an important issue in corpus linguistic research. XML has become a *de facto* standard due to its flexibility and readability as well as the possibility to include annotation and metadata. The Text Encoding Initiative (*TEI Consortium* 2017) provides comprehensive guidelines for storing a variety of text types as XML corpora, and *Hardie* (2014) suggests less 'overwhelming' principles for smaller projects. To allow for exchange and reanalysis, corpus processing should store the corpus in a standard XML format before tailoring the material to specific analytic needs. If conversions are necessary, the Pepper toolkit (*Zipser* et al. 2015) offers tools for converting corpora between a variety of standards. Converting a corpus to an exchangeable format may initially add a step to the workflow, but also means that subsequent operations and analytic steps can easily be repeated for other corpora, and the corpus can be archived and made available for future (re-)analysis and to other researchers.

3. A 'Corpus Workbench' for geographic research

Comprehensive corpora offer new possibilities for inquiry and collaboration. However, few researchers in the field of geography consider themselves experts in computational methods. It is therefore crucial to facilitate access to corpora, as well as analytic tools and methods. Two related open source projects are promising in this respect: The Corpus Workbench (*Evert* and *Hardie* 2011) is a system of data storage and performant query processing. Corpora can be created from XML documents and then searched with a specialised syntax. While the Corpus Workbench comes with an interactive command line tool, CQPweb (*Hardie* 2012) provides a corresponding web interface to perform corpus linguistic analysis functions, such as word frequency counts, collocation analysis and keyword analysis. In CQPweb, sub-corpora can be created based on metadata, for instance, the date of publication, author, or source of a text. Results can be exported as tables for further analysis and visualisation. Analysis functions offered by CQPweb match those offered by other tools, such as AntConc (*Anthony* 2017), WordSmith (*Scott* 2017), or Lexico (lexi-co.com). But in contrast to these tools, the file format and query processor of the underlying Corpus Workbench can handle substantially larger corpora[1] including structural information about texts (e.g. document sub-sections, paragraphs or sentences) and word level annotation (e.g. part-of-speech tags, lemmas or named entities). Whereas CQPweb makes corpus linguistic functionality easily accessible, corpora can also be accessed using 'R' in order to address complex analytic problems (*Desgraupes* and *Loiseau* 2018; *Blaette* 2017b).

The Corpus Workbench and CQPweb are under active development and several universities worldwide host CQPweb, mostly in the context of linguistic research. Our aim is to provide a platform for quantitative text analysis with a particular focus on research in Geography. To this end, we have set up a server at the data centre of the University of Freiburg that is accessible[2] at https://diskurs.geographie.uni-freiburg.de. We currently host an instance of the Polmine GermaParl corpus that includes protocols of debates in the German *Bundestag* since 1998 (*Blaette* 2017a). We prepared several German media corpora, such as Die Zeit (1946-2016), Spiegel Online (1999-present) and tagesschau.de (2015-present), which can be made available if copyright restrictions are observed. All corpora contain lemmas and part-of-speech tags. To facilitate the use of CQPweb, we have published a tutorial that introduces the query syntax and main analytic functions (*Schopper* and *Wiertz* 2017), and plan to extend the available material over time. There are three ways for interested researchers to use the platform or contribute to its further development:

- The CQPweb and available corpora greatly facilitate the use of corpus linguistic methods in the context of **teaching and student project work**.
- Researchers can upload and **use their own corpora** through the interface, and may (but do not have to) share their corpus with others.
- We invite researchers to **collaborate in making more corpora accessible to the research community**. Within the scope of our resources, we may be able to assist in the preparation and processing of sources that are of potential interest to other geographers and social scientists.

We should note that our own contribution is minor considering the efforts made in the various open source projects that we work with and rely on. We nevertheless hope that it helps to open the realm of quantitative text analysis to a wider audience in social science and geography.

Notes

[1] The current limit of 2.1 billion tokens will be lifted further with the transition to a new data model (*Evert* and *Hardie* 2015).

[2] To access corpora, registration is required. Please contact the authors.

References

Anthony, L. 2017: AntConc (Version 3.5.0). – Available online at: http://www.laurenceanthony.net/software/antconc, – accessed 22/1/2018

Ash, J., *R. Kitchin* and *A. Leszczynski* 2018: Digital Turn, Digital Geographies? – Progress in Human Geography **42** (1), – doi: 10.1177/0309132516664800

Blaette, A. 2017a: GermaParl: Corpus of Plenary Protocols of the German Bundestag, TEI Files. – Online available at: https://github.com/PolMine/GermaParlTEI, – accessed 18/12/2017

Blaette, A. 2017b: polmineR (v0.7.5). – Available online at: http://www.github.com/PolMine/polmineR, – accessed 22/1/2018

Bubenhofer, N. 2009: Sprachgebrauchsmuster, Korpuslinguistik als Methode der Diskurs- und Kulturanalyse. – Berlin, Boston, – doi: 10.1515/9783110215854

DeLyser, D. and *D. Sui* 2013: Crossing the Qualitative-Quantitative Divide II: Inventive Approaches to Big Data, Mobile Methods, and Rhythmanalysis. – Progress in Human Geography **37** (2): 293-305, – doi:10.1177/0309132512444063

Desgraupes, B. and *S. Loiseau* 2018: rcqp: Interface to the Corpus Query Protocol. – Available online at: https://cran.r-project.org/web/packages/rcqp/, – accessed 25/3/2018

Dzudzek, I., *G. Glasze*, *A. Mattissek* and *H. Schirmel* 2009: Verfahren der lexikometrischen Analyse von Textkorpora. – In: *Glasze, G.* and *A. Mattissek* (eds.): Handbuch Diskurs und Raum. – Bielefeld: 233-260

Evert, S. and *A. Hardie* 2011: Twenty-First Century Corpus

Workbench: Updating a Query Architecture for the New Millennium. – Proceedings of the Corpus Linguistics 2011 Conference. – University of Birmingham, UK

Evert, S. and *A. Hardie* 2015: Ziggurat: A New Data Model and Indexing Format for Large Annotated Text Corpora. – In: *Banski, P.*, *H. Biber*, *E. Breiteneder*, *M. Kupietz*, *H. Lüngen* and *A. Witt* (eds.): Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3). – Mannheim: 21-27

Glasze, G. 2008. Vorschläge zur Operationalisierung der Diskurstheorie von Laclau und Mouffe in einer Triangulation von lexikometrischen und interpretativen Methoden. – Historical Social Research **33** (1): 185–223

Glasze, G. 2013: Politische Räume: Die Diskursive Konstitution eines geokulturellen Raums – Die Frankophonie. – Bielefeld

Glasze, G. and *F. Weber* 2014: Die Stigmatisierung der Banlieues in Frankreich seit den 1980er Jahren als Verräumlichung und Ethnisierung gesellschaftlicher Krisen. – Europa Regional 2012 (2–3): 63–75

Graham, M. and *T. Shelton* 2013. Geography and the Future of Big Data, Big Data and the Future of Geography. – Dialogues in Human Geography **3** (3): 255–261, – doi: 10.1177/2043820613513121

Gries, S. 2009: Quantitative Corpus Linguistics with R: A Practical Introduction. – New York

Hardie, A. 2012: CQPweb: Combining Power, Flexibility and Usability in a Corpus Analysis Tool. – International Journal of Corpus Linguistics **17** (3): 380-409, – doi: 10.1075/ijcl.17.3.04har

Hardie, A. 2014: Modest XML for Corpora: Not a Standard, but a Suggestion. – ICAME Journal **38** (1): 73–103. doi: 10.2478/icame-2014-0004

Jockers, M.L. 2014: Text Analysis with R for Students of Literature. Quantitative Methods in the Humanities and Social Sciences. – Cham

Linnemann, K. and *P. Reuber* 2015: Der lange Schatten der Moderne: Entwicklungs- und geopolitische Diskurse deutscher Hilfsorganisationen. – Geographische Zeitschrift **103** (1): 1-18

Manning, C.D., *M. Surdeanu*, *J. Bauer*, *J. Finkel*, *S.J. Bethard* and *D. McClosky* 2014: The Stanford CoreNLP Natural Language Processing Toolkit. – Association for Computational Linguistics (ACL) System Demonstrations: 55–60. – Online available at: http://www.aclweb.org/anthology/P/P14/P14-5010, – accessed 18/12/2017

Mattissek, A. 2008: Die neoliberale Stadt: Diskursive Repräsentationen im Stadtmarketing deutscher Grossstädte. – Bielefeld

Mattissek, A. and *G. Glasze* 2016: Discourse Analysis in German-Language Human Geography: Integrating Theory and Method. – Social & Cultural Geography **17** (1): 39–51. doi: 10.1080/14649365.2014.961532

Mattissek, A. and *C. Sturm* 2017: How to Make Them Walk the Talk: Governing the Implementation of Energy and Climate Policies into Local Practices. – Geographica Helvetica **72** (1): 123–135, – doi: 10.5194/gh-72-123-2017

Munzert, S. 2015: Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. – Chichester, West Sussex, United Kingdom

Schmid, H. 1999: Improvements in Part-of-Speech Tagging with an Application to German. – In: *Armstrong, S.*, *K. Church*, *P. Isabelle*, *S. Manzi*, *E. Tzoukermann*, *D. Yarowsky* (eds.): Natural Language Processing Using Very Large Corpora: Text, Speech and Language Technology vol 11, 13-25. – Dordrecht, – doi: 10.1007/978-94-017-2390-9_2.

Schopper, T. and *T. Wiertz* 2017: Korpuslinguistische Analysen mit CQPweb: Eine Einführung für SozialwissenschaftlerInnen. – Online available at: https://human.geographie.uni-freiburg.de/diskurs/doc/cqpweb_einführung_2017-12-01.pdf, – accessed 18/12/2017

Scott, M. 2016: WordSmith Tools version 7, Stroud: Lexical Analysis Software. – Online available at: http://lexically.net/wordsmith/, – accessed 9/3/2018

Silge, J. 2017: Text Mining with R: A Tidy Approach. – Available online at: http://tidytextmining.com, – accessed 18/12/2017

TEI Consortium 2017: TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.2.0. – Available online at: http://www.tei-c.org/Guidelines/P5/, – accessed 18/12/2017

Zipser, F., *T. Krause*, *A. Lüdeling*, *M. Neumann*, *M. Stede* and *A. Zeldes* 2015: ANNIS, SaltNPepper & PAULA: A Multilayer Corpus Infrastructure. – Berlin