

Creating dictionaries for argument identification by reference data

Andreas Niekler¹
aniekler@informatik.uni-leipzig.de

Gregor Wiedemann¹
gregor.wiedemann@uni-leipzig.de

Sebastian Dumm²
sebastian.dumm@hsu-hh.de

Gerhard Heyer¹
heyer@informatik.uni-leipzig.de

¹Universität Leipzig, Germany; ²Helmut-Schmidt-Universität / Universität der Bundeswehr

Introduction

The creation of dictionaries is an important task to conceptualize and operationalize research questions in content analysis (Neuendorf, 2002). One can define concepts for coding operationalized variables in the form of mutual exclusive categories or decide if the content of documents is relevant for coding within the research task by the formalization of meaning through dictionaries (Krippendorff, 2004). Dictionaries are often defined on the basis of a "theory of meaning that reflects a research question or the vocabulary of an academic discipline" (Krippendorff, 2004). Thus, we can think of dictionaries as operationalized representations of historical, sociological, cultural or political theories that are investigated within humanities research. Procedures to create dictionaries automatically allow for more methodical and reproducible research designs especially when dealing with large corpora.

Approach

In contrast to manual dictionary creation from a small set of selected sample documents we present an approach to automatically extract dictionaries from a reference corpus of arbitrary size. Within the "ePol-project" the goal of identifying arguments for a political science research task is approached by creation of two dictionaries. One *semantic dictionary* on the utilization of topic models (Blei/Ng/Jordan, 2003; Teh/Jordan, 2010) to identify thematically relevant documents; and one rather *syntactical dictionary* based on paradigmatic similarity of linguistic markers to identify a high density of argument structures. This poster presents idea, results and an example application of extracted dictionaries for relevancy ranking of retrieval results in large digital document collections.

Semantic dictionaries

Domain experts easily can compile a small reference corpus of paradigmatic documents containing contents of their interest. On this reference corpus we apply a topic model based on the Pitman-Yor Process (Teh, 2006). It employs Poisson instead of Dirichlet distributions which better approximate distributions of natural language data. One of the key properties of topic modeling is the inference of not directly observable variables considered as latent topics. A distribution over these latent topics (classes of co-occurring terms) is allocated to each of the documents within a digital text collection. Another hidden variable describes each of those topics in form of a probability distribution over the vocabulary of the text collection. On the basis of the assumption that all of the topics, extracted in a certain abstraction level controlled by the model parameters, represent the meaning and content of a digital text collection in a compressed form we suggest a partly supervised dictionary extraction process: 1) the set of all resulting topics z is utilized to calculate scores for each word in the vocabulary within the collection. The score for each word is calculated by

$$score(w_n) = \log(F(w_n)) \sum_{k=1}^K p(w_n | z_k),$$

where $p(w_n | z_k)$ is the probability of the n th word in the vocabulary within the k th topic of the model and $F(w_n)$ is the absolute word frequency of the term w_n within the text collection. The idea behind this formula is that terms of high probability within a topic contribute significantly to the meaning of the text collection. At the same time only few terms in a topic have relatively high probabilities. Summing up term probabilities over each topic yields a limited number of the most probable words. Furthermore we take the frequency of words into account because high frequent use of terms and high probability within a topic imply prototypical usage within the texts.

Using topic models further allows for filtering of unwanted semantical structures when creating dictionaries from the collections. In our application we identified a foreign language 'topic' and a topic thematically not related to our research question in the reference texts and could easily exclude them from our k topics before applying the score calculation.

Project information

"ePol - Post-democracy and Neoliberalism"
An eHumanities project funded by the Federal Ministry of Education and Research Germany (05/2012 - 04/2015).

Research institutions:

Helmut-Schmidt-Universität / Universität der Bundeswehr Hamburg
Fakultät für Wirtschafts- und Sozialwissenschaften
Lehrstuhl für Politikwissenschaft, insbesondere Politische Theorie
Prof. Dr. Gary S. Schaal

Universität Leipzig
Institut für Informatik
Lehrstuhl für Automatische Sprachverarbeitung (ASV)
Prof. Dr. Gerhard Heyer

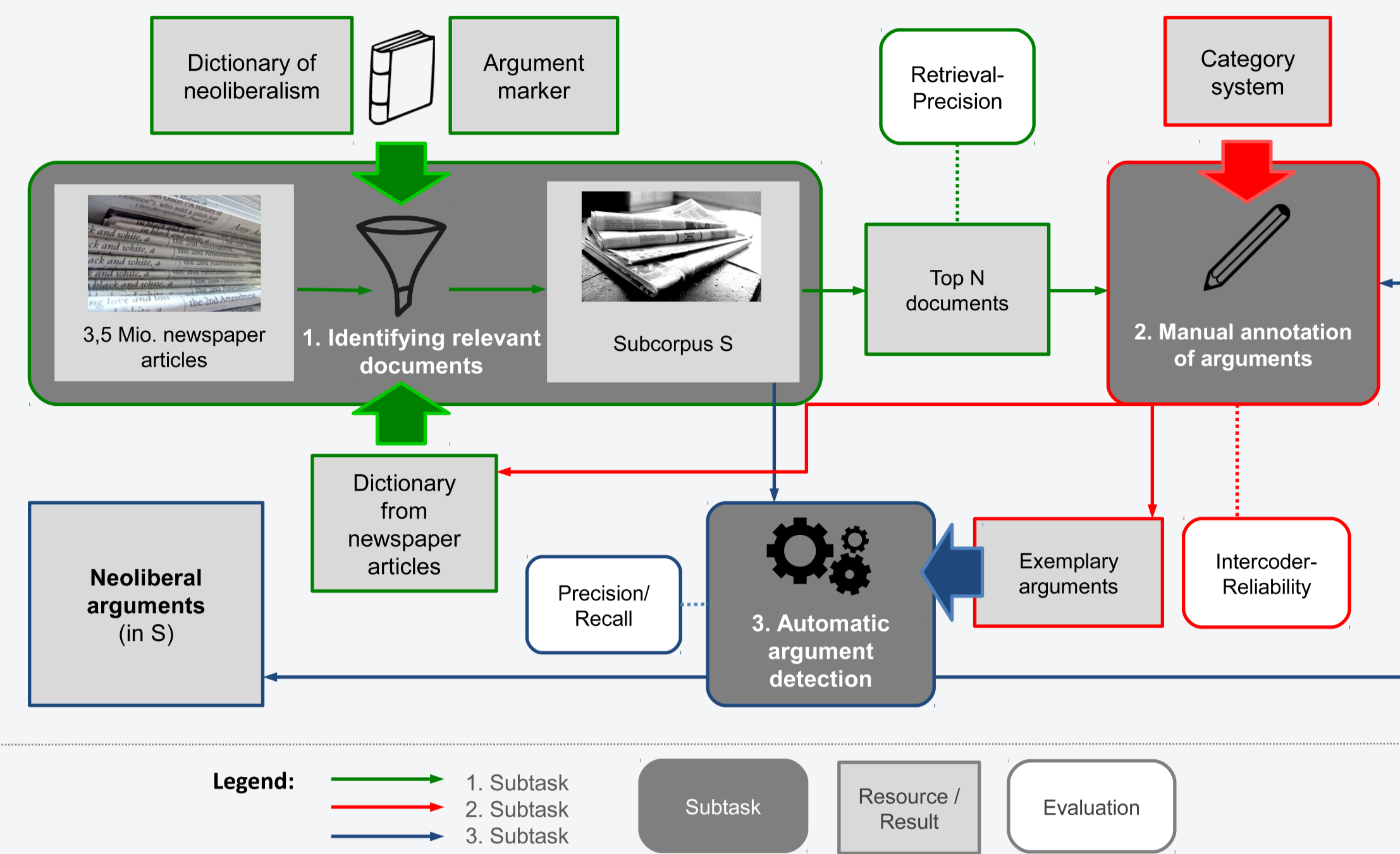


Figure 1: Research design of ePol project shown as a phase model

Syntactic dictionaries via term similarities

Additionally to our dictionary containing semantic information related to theoretical aspects of the political science research task we created a second dictionary of linguistic markers, called *Argumentmarker*, which can be employed to identify argumentative structures (Dumm/Lemke 2013). We took a list of 46 German linguistic markers from another research project on causality and textual coherence (Breindl/Walter, 2009) as a starting point. This list was incrementally extended up to 127 terms by automatically computed synonyms of the markers retrieved from the database of the "Projekt Deutscher Wortschatz" (Quasthoff/Eckart, 2009), a representative corpus of German language.

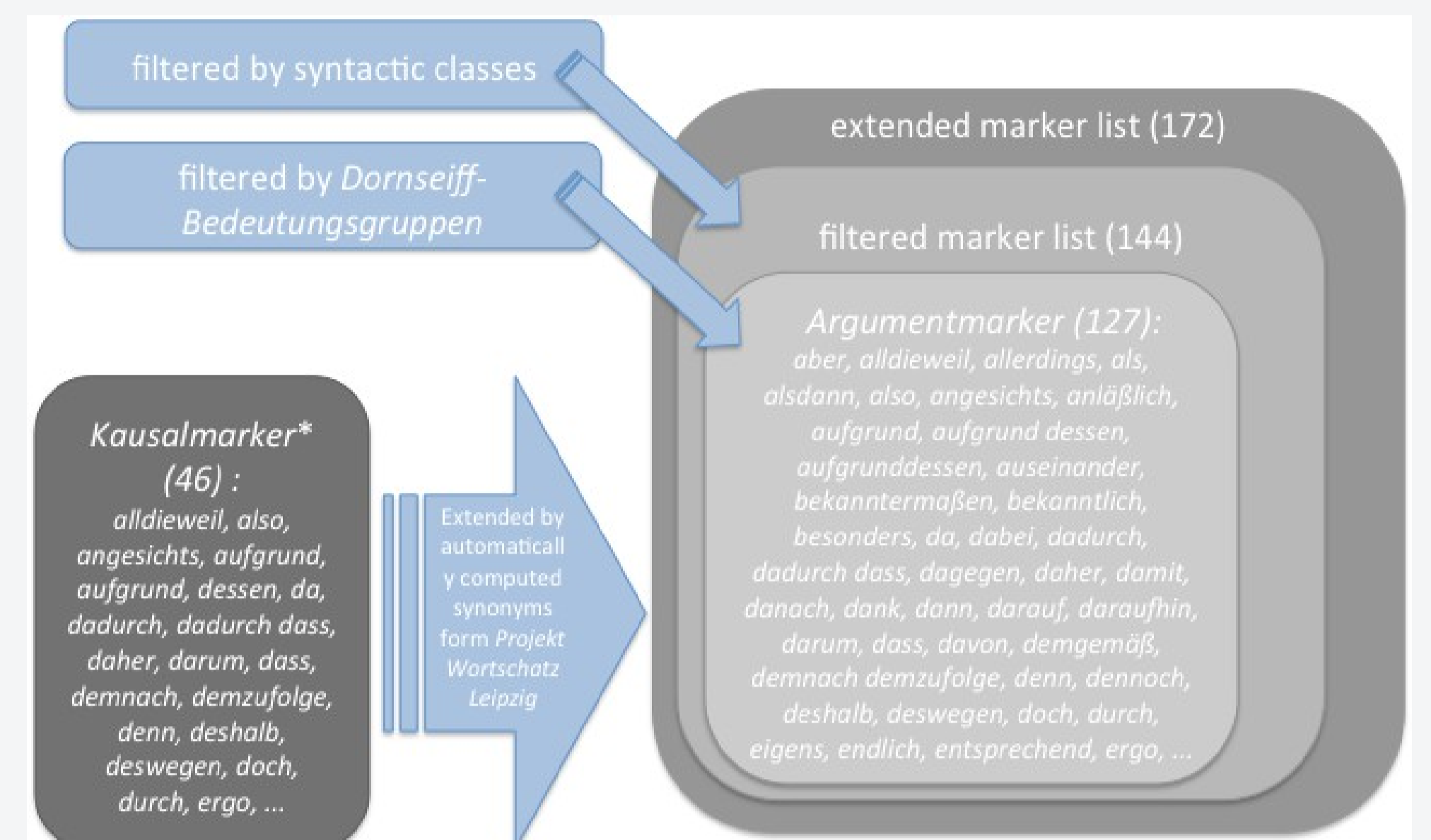


Figure 2: Genesis and filtering process of the Argumentmarker dictionary. The Kausalmarkers* are from Breindl/Walter, 2009.

Application

We applied these dictionaries for retrieval of documents in a large collection of newspaper articles to identify argumentative texts with a certain neoliberal framing. First, a subcorpus S of 10.000 thematically relevant articles is retrieved with a specialized retrieval process (Wiedemann/Niekler 2014) out a corpus of 3.5 million articles using the semantic dictionary as query. These 10.000 documents then are ranked with the *syntactical dictionary* of *Argumentmarker* to retrieve documents with (potentially) high density of argumentations. The best ranked N texts then are subject of a close reading process by political scientists who also utilize the dictionaries for qualitative coding schemes.

score	document length	year	cm	title
347,22	685	1977	58	Pro und kontra Mehrwertsteuer [Pro's and con's of VAT]
321,81	662	1973	45	Ölkrise und Konjunktur [Oil crisis and economy]
290,48	705	1966	36	Energie muß billig sein [Energy has to be cheap]
289,34	687	1977	44	Die Steuern senken [Lower taxes]
287,26	845	1964	54	Korrektur der Einkommensteuer [Correction of VAT]
281,07	687	1971	42	Die Bauern im Nacken [The farmers at the neck]
279,74	884	1965	48	Was ist uns die Mark wert? [What is the „Mark“ worth to us?]
272,75	682	1970	38	Steuern mit der Steuer [Governing with taxes]
264,82	719	1971	36	Ohne Abkühlung keine Stabilität [No stability without slowdown]
262,81	671	1973	39	Das sicherste Mittel [The most secure instrument]
261,33	707	1972	37	Entlastung - wovon? [Relief - of what?]
254,97	676	1979	48	Das Fernsehen und die Angst [Television and fear]
254,93	704	2011	53	Nicht ernst gemeint: die Quote [Quotas not meant serious]
251,53	457	1977	26	Eine Konfliktstrategie der Union [A conflict strategy of the EU]
250,46	638	2010	34	Die neue Auflehnungsbereitschaft [The new willingness for rebellion]

Table 3: Example documents retrieved by a retrieval process based on a semi-automatic extracted dictionary in combination with a causal marker dictionary. The sorted lists represents relevant documents from large document collections with a high argumentation density.

References

- Alsumait, L., Barabá, D., Gentile, J., & Domeniconi, C. (2009). Topic Significance Ranking of LDA Generative Models. In Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part I (S. 67-82). Berlin, Heidelberg: Springer-Verlag.
- Breindl, Eva / Walter, Maik (2009): Der Ausdruck von Kausalität in Deutschen. Amades - Arbeitspapiere zur deutschen Sprache, Mannheim.
- Blei, D.M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. The Journal of Machine Learning Research, 3, 993-1022.
- Dumm, S., Lemke, M. (2013). Argumentmarker. Definition, Generierung und Anwendung im Rahmen eines semi-automatischen Dokument-Retrieval-Verfahrens. Hamburg. Discussion Paper ePol.
- Krippendorff, K. (2004). Content analysis: an introduction to its methodology (2nd ed.). Thousand Oaks Calif.: Sage.
- Neuendorf, K. A. (2002). The content analysis guidebook. Thousand Oaks, Calif: Sage Publications.
- Niekler, A., & Jähnichen, P. (2012). Matching Results of Latent Dirichlet Allocation for Text. In Proceedings of ICCM 2012, 11th International Conference on Cognitive Modeling (S. 317-322). Universitätsverlag der TU Berlin.
- Quasthoff, Uwe / Eckart, Thomas (2009): Corpus Building Process of the Project "Deutscher Wortschatz". In: Linguistic Processing Pipelines Workshop at GSCS 2009.
- Teh, Y. W., & Jordan, M. I. (2010). Hierarchical Bayesian Nonparametric Models with Applications. In N. Hjort, C. Holmes, P. Müller, & S. Walker (Hrsg.), Bayesian Nonparametrics: Principles and Practice. Cambridge University Press.
- Teh, Yee Whye. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In Proceedings of the 21st International Conference on Computational Linguistics, 985-992.
- Wiedemann, G. / Niekler, A. (2014): Document Retrieval for Large Scale Content Analysis using Contextualized Dictionaries, Schriftenreihe des Verbundprojekts Postdemokratie und Neoliberalismus. Discussion Paper Nr. 2.

Term	Weigth
jahr NN [year]	0.37938
gross ADJA [big]	0.368788
einkomm NN [income]	0.353946
preis NN [price]	0.344253
gut NN [goods]	0.293837
polit ADJA [political]	0.289046
zeit NN [time]	0.27682
hoh ADJA [high]	0.240263
kost NN [cost]	0.231548
regeln NN [rules]	0.221033
mensch NN [human]	0.217523
offent ADJA [public]	0.215913
person NN [person]	0.212456
regier NN [government]	0.210173
wert NN [value]	0.208864
inflation NN [inflation]	0.201939
anlys NN [analysis]	0.200244
bestimmt ADJA [certain]	0.199011
allgemein ADJA [common]	0.198007

Table 2: Top 20 terms extracted via topic model. Not related topics were excluded from the creation process.