Extending Corpus-Based Discourse Analysis for Exploring Japanese Social Media

Philipp Heinrich[1] **and Fabian Schäfer**[2]
Friedrich-Alexander-University Erlangen-Nuremberg
[1]Chair of Computational Corpus Linguistics, Bismarckstr. 6, 91054 Erlangen
[2]Chair of Japanese Studies, Artilleriestr. 70, 91052 Erlangen
{philipp.heinrich, fabian.schaefer}@fau.de

Abstract

The present paper presents and applies novel techniques for corpus-based discourse analysis in order to analyze Japanese social media data. Firstly, we visualize collocational profiles of topic nodes (e. g. lexical items or corpus queries) by projecting word embeddings onto a two-dimensional semantically structured plane, where the size of the displayed items represent their association strength to the topic node. Secondly, we generalize collocations in two ways: by looking at collocates of discourses (which are created by combining several sets of lexical items), and by looking at second-order topic collocates of lexical items. We perform a discourse analysis in a corpus of more than 400,000,000 tweets produced before and in the aftermath of the Fukushima Daiichi nuclear disaster with a focus on discussions about (phasing out) nuclear energy.

Keywords: Corpus-Based Discourse Analysis, Social Media, Fukushima, Visualization, Collocations

1. Introduction

The Fukushima Daiichi nuclear disaster in March 2011 has led to discussions about "energy transition" and the phasing out of nuclear energy throughout the world; a phenomenon called the Fukushima Effect (Arlt and Wolling, 2016). Different political camps take different stances towards the topic and unsurprisingly it has been widely discussed in the run-up to political elections. Previous research (Abe, 2015; Yoshino, 2013) e. g. showed that Japanese newspapers can be categorized into pro-nuclear ones on the one hand (such as the conservative newspapers Sankei and Yomiuri) and anti-nuclear ones on the other (such as the left-leaning Asahi). At the same time, there is no systematic corpus-based study of the impact that the Fukushima incident had on social media. The paper at hand thus provides an analysis of Japanese social media data (namely Twitter)[1] shortly before and in the aftermath of 3/11.

The results presented here are part of a broader line of research in which we analyze the *transnational algorithmic public sphere* (Heinrich et al., 2018); one aspect being communication that is mediated through algorithms and that spreads through the Internet. As such, we are concerned with the tempo-spatial dissemination of discourses across social media. The Fukushima incident can be seen as a particularly insightful starting point, since the shift in attitudes and opinions towards nuclear energy are to be observed on a global scale.

Methodologically, we build on corpus-based discourse analysis (CDA) (Baker, 2006), but are extending it by two novel techniques first sketched in Heinrich et al. (2018) and refined here. Firstly, the bottleneck of CDA is the amount of hermeneutic interpretation necessary to analyze the whole corpus, which is why we actively develop and refine visualization techniques for keyword and collocation analyses (which are at the heart of CDA). Secondly we triangulate the semantics of discourse nodes and their collocates by looking at higher-order collocates. This way we can explore the interplay of discourses, such as the phase-out of nuclear energy in the context of elections. Our long-term goal is the development of methods which enable hermeneutic researchers to analyze and qualitatively interpret huge amounts of textual data without excessive cherry-picking.

[1]Twitter has several advanteages over other social media networks, including accessibility, metadata availability, sample size, and brevity (Mejova et al., 2015; Burghardt, 2015).

2. Methodology

2.1. Data and Preprocessing

We are using a sample of 411,452,027 Japanese original tweets (excluding retweets) whose identifiers have been collected via the public streaming API of Twitter[2] between 2010 and 2015. We downloaded the whole tweets (text and meta data) using the tweet identifiers via the REST API in 2016. Since tweets which have been deleted by the user or which were written by users who have been suspended by Twitter cannot be retrieved via the REST API, our analysis is not based on unbiased data – tweets which represent an unpopular opinion might be erased more often than others, and social bots might get identified by Twitter and their tweets will subsequently not be part of our collection. Nevertheless, we were able to download more than 80% of the tweets for which we had collected identifiers. Including further data taken from the tweets downloaded via the REST API (e. g. the originals belonging to retweets in our data set) results in the number of tweets mentioned above.

Moreover, we are actually interested in *removing* data produced by social bots in order to reduce the amount of noise omnipresent in social media data (Schäfer et al., 2017). We thus remove any near-duplicates from the data set, which reduces the number of tweets by another 19.5%; the final corpus consists of 331,306,663 original postings between 2007 and 2015.

It is worth mentioning that the deduplication process has little to no effect on our collocation analyses. Looking at

[2]`https://developer.twitter.com/en/docs.html`

the topic node *Fukushima* around 3/11, e. g., we see that (1) there is a slightly higher node frequency in the deduplicated data set (this means that bots have been comparatively more active in other parts of the data set) and (2) hashtags such as #save_fukushima are a bit less salient in the deduplicated data – probably due to the fact that automated, semi-automated, or simply copied data is more prone to contain pre-fabricated items such as hashtags.

The whole data is processed using MeCab[3] using the ipadic-neologd dictionary (Sato et al., 2017), which splits Japanese texts into short-unit morphemes. A custom stopwords lists is used to filter out grammatical particals, punctuation marks, etc.

2.2. Corpus-Based Discourse Analysis

CDA basically boils down to the aggregation and subsequent deconstruction of concordance lines (Baker, 2006; Baker et al., 2008). The categories in which the textual data is divided in a CDA have to be made up by the hermeneutic researcher while dealing with the data. Discourses are formed around lexical items, and thus discourse analysis starts with a corpus query representing the topic node, e. g. a regular expression such as (原子*)|(原発) representing the topic "nuclear energy". For Twitter data, we assume that all tweets matched by the corpus query are part of the discourse.

Collocations of the topic node are considered possible attitudes or stances towards the topic; a prototypical example for such a pair of topic and attitude would be "refugees as victims" (Baker, 2006, 86). In the study at hand, we operationalize collocates via textual co-occurrence, i. e. lexical items that co-occur within the same tweets as the topic are the basis for our *collocation analysis* (Evert, 2008). Let w_1 denote the topic node. We use the following contingency table as the basis for calculating statistical association between w_1 and all collocation candidates w_2 across all tweets t:

	$w_2 \in t$	$w_2 \notin t$	
$w_1 \in t$	O_{11}	O_{12}	$= R_1$
$w_1 \notin t$	O_{21}	O_{22}	$= R_2$
	$= C_1$	$= C_2$	$= N$

In the paper at hand, we use Log-Likelihood (LL) as a statistical association measure for retrieving collocates, with

$$LL(O_{11}, O_{12}, O_{21}, O_{22}) = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}},$$

where $E_{ij} = R_i C_j / N$ denotes the expected absolute co-occurrence frequency in the case of independence of occurrences. We typically analyze the 50–100 top collocates when performing a discourse analysis.

It is worth noting that *keyword analyses*, which compare frequencies of two corpora with one another in order to determine which words are salient for the respective corpus, are very similar to collocation analyses. In fact, a textual collocation analysis is quasi-identical to a keyword analysis in which the subcorpus is specified by a corpus query specifying the lexical item(s) for which collocates should be retrieved; the only difference being that the topic node is included in keyword analyses and is explicitly excluded from collocation analyses. It is thus unsurprising that the same statistical association measures are used both for retrieving keywords as well as collocates. Hence we will use the words *keyword analysis* and *collocation analysis* interchangeably.

With our focus being the methodological advancement of the field and for reasons of brevity, we will not present at *concordances*, i. e. actual tweets containing topic and collocate, when interpreting the data. It is important to emphasize, however, that a qualitative corpus interpretation should never lose sight of the actual evidence provided by the corpus.

2.3. Higher-Order Collocates

In order to triangulate the semantics of discourses[4] identified by our technique, we generalize the operationalization of collocates within CDA in two ways. Firstly, we look at discourse collocates. The basic idea here is to analyze the cooccurrence frequencies of the discourse occurrences. For each lexical item w, we thus compare the frequencies of w in all tweets which have been identified to be part of the discourse (such as containing both 日本 (*Japan*) and being matched by (原子*)|(原発) (*nuclear energy*) with the marginal frequencies of w. The top collocates represent lexical items that play a role in the discourse at hand (e. g. Japan's use of nuclear energy).

Furthermore, we look at topic-based second-order collocates; the idea here is to analyze the cooccurrences of one set of lexical items c (e. g. 日本) in tweets that are about a certain topic t (e. g. tweets matched by (原子*)|(原発)). For each lexical item w we thus compare the cooccurrence frequencies of w with c among tweets that contain topic t with the marginal frequencies of w among all tweets containing topic t. This way we can find out what collocates of t (*nuclear energy*) are particularly important for c (*Japan*).

2.4. Visualization Techniques

Since CDA means interpreting n-best lists of collocates (or keywords), with n potentially becoming confusingly large (larger than, say, 100), we facilitate the interpretation of these lists by a custom visualization technique. Firstly, we create high-dimensional word embeddings (Mikolov et al., 2013) for the specific linguistic register at hand (Japanese social media data), using an unrelated sample of 133,526,833 deduplicated and preprocessed Japanese tweets collected between February 2017 and June 2018 via the Streaming API. We then project the embeddings of an n-best list into a two-dimensional semantically structured space using t-distributed stochastic neighbour embedding

[3] https://taku910.github.io/mecab/

[4] For the study at hand, we follow traditional CDA terminology and refer to pairings of topics and collocates as discourses. However, our methodology should be seen as an intermediary step towards identifying discourses. Pairings of topics with collocate group(s) should rather be viewed as discourse *building blocks* which have to be combined with one another in order to identify discourses in the Foucauldian sense.

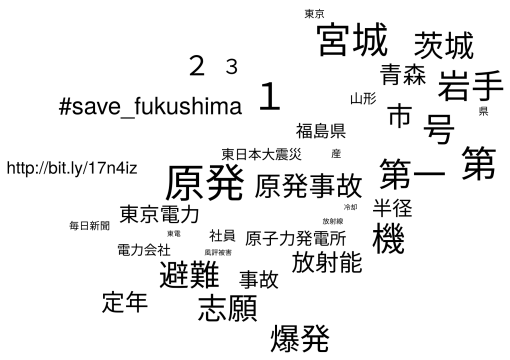

Figure 1: Visualiation of the top-50 collocates of (the discourse around) the topic node 福島 (*Fukushima*) in the aftermath of 3/11. The topic node appears in 29,425 out of 5,744,937 tweets observed between March 12 and March 19, 2011, amounting to roughly 5,122 tweets per million.

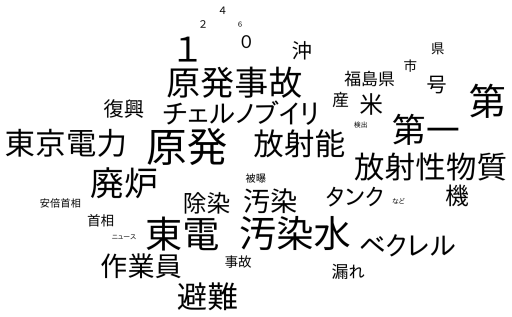

Figure 2: Visualization of the top-50 collocates of 福島 (*Fukushima*) in the end of 2013.

(t-SNE) (van der Maaten and Hinton, 2008).[5] Semantically similar lexical items are expected to appear in the vicinity of one another since they have similar embeddings.

For an outcome of our procedure, see Figure 1. Once can clearly see that there are regions in the plot that contain places which are affected by the accident on the one hand, and e. g. digits (indicating parts of the Fukushima Daiichi Nuclear Power Plant) somewhere else – that is to say, the plane is indeed semantically structured. Furthermore, the size of the displayed lexical items reflects the collocational strength (a rescaled value of the association measure) of the item towards the topic at hand. In Figure 1, examples for important collocates of Fukushima are expectedly 原発 (*nuclear power* (*plant*)), 原発事故 (*nuclear accident*), and the hashtag #save_fukushima; a less important one is 東電 (*TEPCO*).

Note that the procedure is stochastic, i. e. for each run of t-SNE the result will be different. In order to directly compare several of these plots with one another, the embeddings of all lexical items occurring in any of the plots thus have to be projected in a single t-SNE run, a point to keep in mind when comparing Figure 1 with Figure 2. Note that it is not feasible to run t-SNE on *all* embeddings beforehand. With the number of encountered token types easily exceeding millions – especially when dealing with social media data – the projection is computationally too costly.

3. Empirical Analysis

We observe a selection of topic nodes and their collocates from the beginning of 2010 until 2014. For an estimate of their frequency in our corpus (and thus on Twitter), see Figure 3. The estimates are based on at least 5,000,000 tweets with an observed topic frequency of at least 5,000 tweets, i. e. we create our windows of observation subject to the amount of data we observe over time. Note the logarithmic y-axis, seemingly small differences amount to large differences in the absolute amount of tweets.

3.1. Fukushima

福島 (*Fukushima*, green line) has always been a topic on Twitter. However, we observe an unsurprisingly huge increase in the number of tweets between March 11 and March 16, 2011 (from a stable 250 to more than 5,000 tweets per million). The amount of tweets containing the topic node decreases very slowly in the aftermath and remains above pre-3/11 levels until the end of our observation period.

Most of its important collocates during this time period are lexical items referring to the accident (原発, 原発事故) and the hashtag #save_fukushima, but also the electric utility holding company 東電 (*TEPCO*) in charge of the power plant, cf. Figure 1.

When comparing this collocational profile of the node with the one in the end of 2013, e. g. between September 10 until November 16, 2013 (when the frequency has fallen to around 500 tweets per million) many of these collocates remain, see Figure 2. Additionally, the focus shifts to political actors 安倍首相 (*Prime Minister Shinzō Abe*) and the results of and measures taken due to the radioactive accident – 除染 (*decontamination*), 汚染水 (*contaminated water*), 放射能 (*radioactivity*) – appear in the discourse.

3.2. Elections

Expectedly, there are huge peaks in the number of tweets about 選挙 (*elections*, purple line), at 3/11 and at dates which coincide e. g. with the elections of Tokyo's governor after resignation of 石原 (*Shintaro Ishihara*). In fact, 石原 is one of the top collocates of 選挙 in the period between April 04 and April 10, 2011, where its frequency peaks with more than 3,000 tweets per million. Other importand collocates are 結果 (*results*), 都知事選 (*gubernatorial election*), and 候補(者) (*candidate*, *candidacy*), cf. Figure 4.

In the end of 2012 (December 05, 2012 till January 01, 2013), elections are discussed frequently again, with more than 2,200 tweets per million. The most important collocates have shifted towards 自民 (*Liberal Democratic Party;*

[5] Lexical items for which there are no embeddings available (since they have not been encountered frequently enough in the data set used for creating the embeddings) are initialized at random locations of the plane.

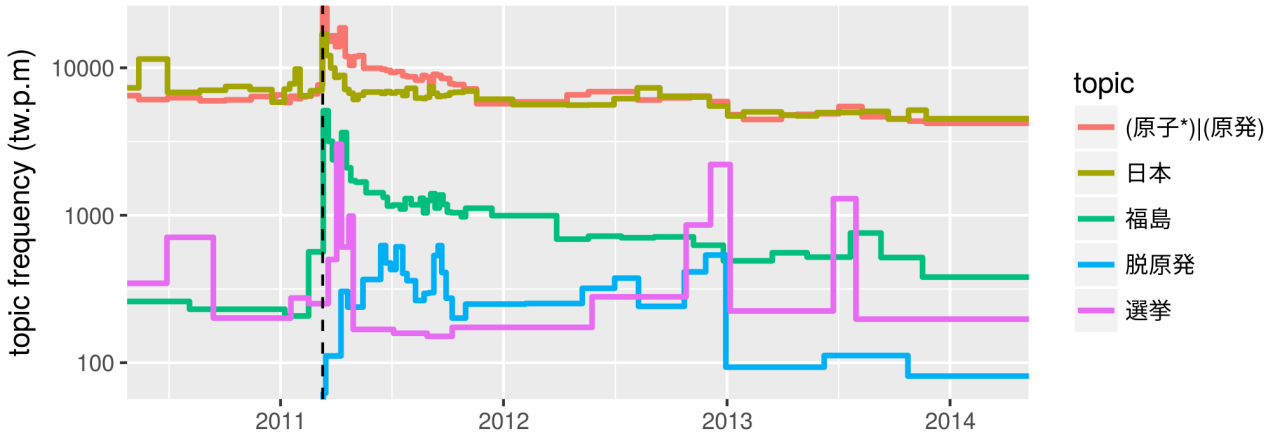

Figure 3: Frequencies (in tweets per million) of selected topics during the observation period on a logarithmic scale. The dashed line represents March 11, 2011. All observed frequencies peak at or shortly after 3/11.

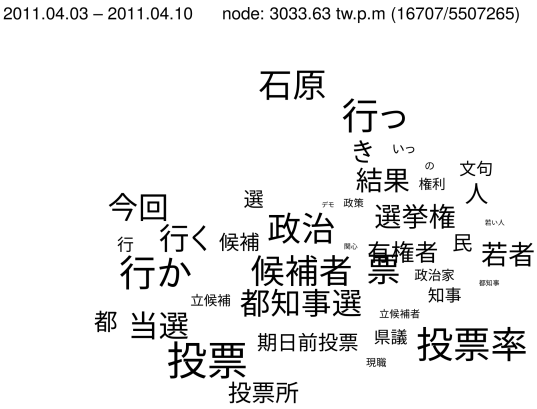

Figure 4: Top-50 collocates of 選挙 (*elections*) in the aftermatch of 3/11.

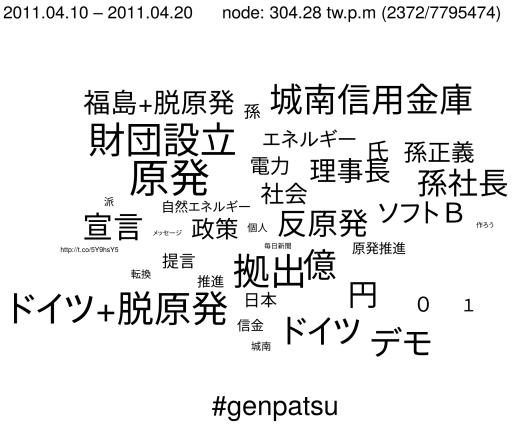

Figure 5: Top-50 collocates of 脱原発 (*nuclear phase-out*) shortly after 3/11.

LDP), and nuclear power (plants) (原発) have taken an important role in the discussion. These collocates remain in the collocational profile until the end of the observation period, whereas the political actors appearing in the profile naturally change: in the last peak in our data (1,300 tweets per million between June 24 and July 30, 2013), actor and anti-nuclear politician 山本太郎 (*Tarō Yamamoto*) has displaced 石原.

3.3. Nuclear Phase-Out

Discussions about phasing out nuclear energy (脱原発, blue line) enter the debate only a couple of weeks after 3/11; the term is literally non-existent on Twitter before 3/11, quickly rising to absolute frequencies of more than 600 tweets per million in July 2011, and decreasing to frequencies of less than 100 tweets per million after 2012.[6] Indeed, whether or not to "break with nuclear energy" is a discussion led elsewhere, e. g. in ドイツ (*Germany*), which appears as an important collocate next to 日本 (*Japan*), cf. Figure 5. Further collocates of 脱原発 comprise 福島 (*Fukushima*), 原発 (*nuclear power plant*), and デモ (*demonstration*). Discussions about phasing out nuclear energy peak another time in the end of 2012 (more than 500 tweets per million between November 27 until December 29), in which political actors such as 未来の党 (the *Tomorrow Party of Japan*) and 山本太郎 (*Tarō Yamamoto*) appear.

3.4. Japan and Nuclear Energy

The more general topic nodes 日本 (*Japan*) and (原子*)|(原発) (*nuclear energy*) are both more frequent (with frequencies of more than 5,000 tweets per million throughout the observation period and temporarily more than 10,000 tw.p.m.) and are selected for further semantic triangulation in the present subsection. Figures 7 to 11 are plots of the same t-SNE run and are thus directly comparable with one another.

Before 3/11, important collocates of 日本 (*Japan*) involve 語 (referring to the *Japanese language*), and countries such as China and the USA, cf. Figure 6. In the aftermath of 3/11, collocates related to the Fukushima Daiichi nuclear

[6] Since 脱原発 appears so infrequently, it is the only node whose relative frequency we measure in variable windows of at least 2,000 absolute topic occurrences.

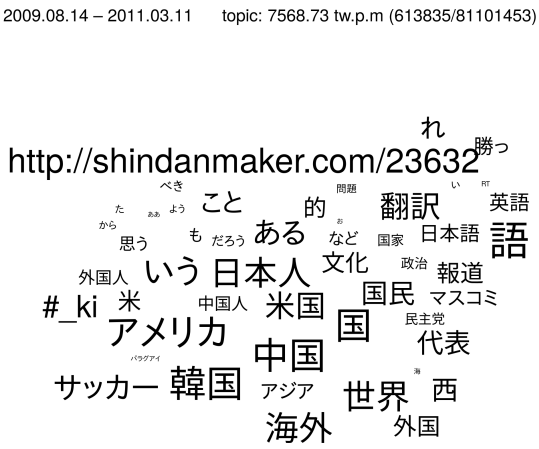

Figure 6: Collocates of 日本 (*Japan*) before 3/11.

Figure 7: Collocates of 日本 (*Japan*) in the aftermath of 3/11.

Figure 8: Collocates of 日本 (*Japan*) between 2012 and 2015.

Figure 9: Discourse collocates of 日本 (*Japan*) and (原子*)|(原発) (*nuclear energy*) in the aftermath of 3/11.

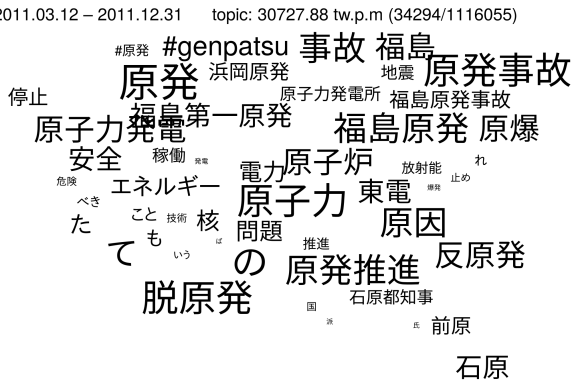

Figure 10: Second-order collocates of 日本 (*Japan*) in tweets matching (原子*)|(原発) (*nuclear energy*) in the aftermath of 3/11.

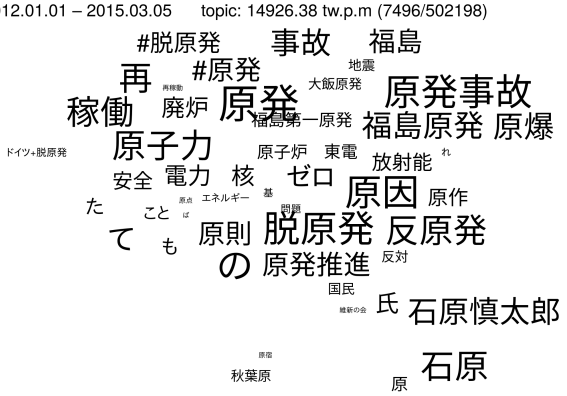

Figure 11: Second-order collocates of 日本 (*Japan*) in tweets matching (原子*)|(原発) (*nuclear energy*) between 2012 and 2015.

disaster enter the semantic profile of 日本 (*Japan*) – such as 地震 (*earthquake*), 復興 (*reconstruction*), 原発 (*nuclear power plant*), and 赤十字社 (*red cross*), cf. Figure 7. Additionally, Germany becomes more important, presumably due to its importance in the political discussions about breaking with nuclear energy. Interestingly, 原発 (*nuclear power plant*) even remains an important collocate after 2012 (cf. Figure 8).

Discussions about (原子*)|(原発) (*nuclear energy*) obviously peak in the aftermath of 3/11. Discourse collocates of 日本 (*Japan*) and nuclear energy in the aftermath of 3/11 are 福島 (*Fukushima*) and 原発事故 (*nuclear accident*), cf. Figure 9. Additionally 脱原発 (*nuclear phase-out*) as well as ドイツ (*Germany*) are important aspects of discussions about nuclear energy in Japan.

Figures 10 and 11 show second-order collocates of 日本 (*Japan*) in the context of nuclear energy in the aftermath of 3/11 and after 2012, respectively, it becomes obvious that the accident in Fukushima has had a lasting effect on more global discourses: next to political actors we can now find items such as 脱原発 (*nuclear phase-out*). Another phenomenon we can observe is that the actual lexical item 脱原発 (*nuclear phase-out*) has been replaced by the hashtag #脱原発.

4. Conclusion

We have presented a corpus linguistic analysis of Japanese social media data in the aftermath of 3/11. It is noteworthy that for Twitter data, textual cooccurrences seem to be an easy and suitable basis for calculating collocations. Further studies might however experiment with other association measures for retrieving collocates, especially when interested in less salient items. On a qualitative empirical level, the paper showed that the nuclear phase-out debate entered Japanese Twitter only several weeks after 3/11. Moreover, the salience of discussions about phasing out nuclear energy and about nuclear energy in general is quite volatile and correlates i. a. with elections. Last but not least, we showed how particular parts of the nuclear energy discussion entered the collocational profile of the very general discourse around Japan.

Methodologically, we presented two contributions to CDA: Firstly, we developed a powerful visualization technique for discourse profiles. We believe that visualization is important since there is a good case to believe that discourse analysis will not be possible to do without hermeneutic (human) interpretation any time soon (if at all). Facilitating this task thus has to be at the heart of the methodological development. Secondly, we presented second-order topic collocates and discourse collocates, which enable the hermeneutic interpreter to dig deeper into the semantics of discourses. However, we encourage the reader to think of topic nodes and collocate (groups) as building blocks of discourses: it is only after combining several building blocks (e. g. *Japan*, *nuclear phase-out*, and *elections*) that we can hope to identify discourses in the narrow sense.

5. Acknowledgements

This work was supported by the *Emerging Fields Initiative* (EFI) of Friedrich-Alexander-Universität Erlangen-Nürnberg (project title: Exploring the 'Fukushima Effect' (EFE)).

6. Bibliographical References

Abe, Y. (2015). The nuclear power debate after Fukushima: a text-mining analysis of Japanese newspapers. *Contemporary Japan*, 27(2), January.

Arlt, D. and Wolling, J. (2016). Fukushima effects in Germany? Changes in media coverage and public opinion on nuclear power. *Public Understanding of Science*, 25(7):842–857, October.

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyżanowski, M., McEnery, T., and Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, 19(3):273–306.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. Continuum, London.

Burghardt, M. (2015). Introduction to Tools and Methods for the Analysis of Twitter Data. *10plus1: Living Linguistics*, 1.

Evert, S. (2008). Corpora and collocations. In Anke Lüdeling et al., editors, *Corpus Linguistics. An International Handbook*, chapter 58. Mouton de Gruyter, Berlin.

Heinrich, P., Adrian, C., Kalashnikova, O., Schäfer, F., and Evert, S. (2018). A Transnational Analysis of News and Tweets about Nuclear Phase-Out in the Aftermath of the Fukushima Incident. In Andreas Witt, et al., editors, *Proceedings of the LREC 2018 "Workshop on Computational Impact Detection from Text Data"*, Paris. ELRA.

Yelena Mejova, et al., editors. (2015). *Twitter: A Digital Socioscope*. Cambridge University Press, New York.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Sato, T., Hashimoto, T., and Okumura, M. (2017). Implementation of a word segmentation dictionary called mecab-ipadic-neologd and study on how to use it effectively for information retrieval (in japanese). In *Proceedings of the Twenty-three Annual Meeting of the Association for Natural Language Processing*, pages NLP2017–B6–1. The Association for Natural Language Processing.

Schäfer, F., Evert, S., and Heinrich, P. (2017). Japan's 2014 General Election: Political Bots, Right-Wing Internet Activism and PM Abe Shinzō's Hidden Nationalist Agenda. *Big Data*, 5:1 – 16.

van der Maaten, L. and Hinton, G. (2008). Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Yoshino, Y. (2013). Datsugenpatsu, hangenpatsuk ōdō ni kansuru shinbunkiji no sōi: Asahi shinbun to yomiuri shinbun (The Differences between Newspaper Articles on the Anti-Nuclear Power Movement. Asahi Shinbun and Yomiuri Shinbun). *Chikushi Jogakuen University*, 8:89–100.