Extending Corpus-Based Discourse Analysis for Exploring Japanese Social Media

Philipp Heinrich[1] and Fabian Schäfer[2]
[1] *Chair of Computational Corpus Linguistics*, [2] *Chair of Japanese Studies*
Friedrich-Alexander University of Erlangen-Nuremberg
September 17, 2018

Introduction

Background

- Exploring the *Fukushima Effect*
  - identification and analysis of the tempo-spatial propagation of **discourses** in the **transnational algorithmic public sphere**
  - case study: Fukushima Effect (cf. Gono'i, 2015)
  - data: mass and social media (German, Japanese)
    - ☞ **Japanese Twitter**
  - `www.linguistik.fau.de/projects/efe/`
  - funded by the **Emerging Fields Initiative** of FAU
- Team:
  - **Chair of Computational Corpus Linguistics**
    Prof. Dr. Stefan Evert, Philipp Heinrich
  - **Chair of Japanese Studies**
    Prof. Dr. Fabian Schäfer, Olena Kalashnikova
  - **Chair of Communication Science**
    Prof. Dr. Christina Holtz-Bacha, Christoph Adrian
  - **Chair of Visual Computing**
    Prof. Dr.-Ing. Marc Stamminger, Jonas Müller

Research Focus

- methodological foundation: Corpus-Based Discourse Analysis (CDA)
- development of novel techniques (Mixed-Methods Discourse Analysis, MMDA):
  - visualization
  - higher-order collocates
- ultimate goal: assist hermeneutic researchers in interpreting huge amounts of textual data without excessive cherry-picking
- lexical nodes in the case study here:
  - 福島 (Fukushima)
  - 選挙 (elections)
  - 脱原発 (nuclear phase-out)
  - 日本 (Japan) + (原子*)|(原発) (nuclear energy)

☞ focus on methodology

Research Focus

- methodological foundation: Corpus-Based Discourse Analysis (CDA)
- development of novel techniques (Mixed-Methods Discourse Analysis, MMDA):
  - visualization
  - higher-order collocates
- ultimate goal: assist hermeneutic researchers in interpreting huge amounts of textual data without excessive cherry-picking
- lexical nodes in the case study here:
  - 福島 (Fukushima)
  - 選挙 (elections)
  - 脱原発 (nuclear phase-out)
  - 日本 (Japan) + (原子*)|(原発) (nuclear energy)

☞ focus on methodology

Methodology

Corpora – mass media

Frankfurter Allgemeine Zeitung (2011–2014)

- statistics:
  - 306,580 articles, 1,656,372 paragraphs
  - 145,055,523 tokens (1,981,726 types)
- linguistic annotation:
  - TreeTagger (tokenization, POS-tagging, lemmatization)

Yomiuri Shimbun (2011–2015)

- statistics:
  - 1,688,435 articles, 12,757,433 paragraphs
  - 580,518,367 tokens (392,971 types)
- linguistic annotation:
  - MeCab (SUWs)

Corpora – social media (Twitter)

German Twitter

- 10,266,835 original posts
- linguistic annotation:
  - tokenization: SoMaJo (Proisl and Uhrig, 2016)
  - POS-tagging: SoMeWeTa (Proisl, 2018)
  - lemmatization: work in progress

Japanese Twitter

- 411,452,027 original posts
- linguistic annotation:
  - MeCab + special dictionary: ipadic-neologd (Sato et al., 2017)

+ removal of noise: approximately 20% (Schäfer et al., 2017)

Corpora – social media (Twitter)

German Twitter

- 10,266,835 original posts
- linguistic annotation:
  - tokenization: SoMaJo (Proisl and Uhrig, 2016)
  - POS-tagging: SoMeWeTa (Proisl, 2018)
  - lemmatization: work in progress

Japanese Twitter

- 411,452,027 original posts
- linguistic annotation:
  - MeCab + special dictionary: ipadic-neologd (Sato et al., 2017)

  + removal of noise: approximately 20% (Schäfer et al., 2017)

Corpus-Based Discourse Analysis (CDA)

- CDA means analyzing and deconstructing concordance lines (Baker, 2006)
  - concordances are the essence of discourses
- finding **discourses**: **nodes** + **attitudes**
  - (topic) nodes: defined by *keywords* or (more generally) *corpus queries*
  - attitudes: *collocates* that are retrieved by statistical methods
- examples
  - “refugees as victims” (Baker, 2006)
  - “Fukushima as worst case scenario”

in practice:

- look at (*n* best) collocates of topic node
- make up categories on the fly
- categorize manually

Corpus-Based Discourse Analysis (CDA)

- CDA means analyzing and deconstructing concordance lines (Baker, 2006)
  - concordances are the essence of discourses
- finding **discourses**: **nodes** + **attitudes**
  - (topic) nodes: defined by *keywords* or (more generally) *corpus queries*
  - attitudes: *collocates* that are retrieved by statistical methods
- examples
  - “refugees as victims” (Baker, 2006)
  - “Fukushima as worst case scenario”

in practice:

- look at (*n* best) collocates of topic node
- make up categories on the fly
- categorize manually

Collocates and Keywords

keywords

- given two frequency lists of lexical items
- perform statistical tests on frequency litss
  - always *viz.* reference corpus
  - measures: log-likelihood, log-ratio, frequency filter

collocates

- given a definition of a subcorpus
- rate lexical items according to association strength
  - windows vs. segments (**textual co-occurrence**)
  - association measures: see above

From Textual Co-Occurrences to Collocates

- contingency table (cf. Evert, 2008)

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

- calculate expected frequencies subject to independence of co-occurrences (E_{ij})
- apply association measure

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

Extension: Higher-Order Collocates

1. discourse collocates
   - straightforward generalization with respect to textual co-occurrence
   - look at co-occurrence frequencies of tweets that were identified to be part of the discourse at hand (topic + attitude)
   - collocates represent lexical items that play a role in the **discourse**
2. second-order topic-collocates (or attitude-collocates)
   - look at co-occurrence frequencies of one set of lexical items *c* in tweets that are about a certain topic *t*
   - collocates of *c* that are particulary important for *t*

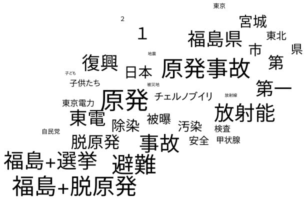
Extension: Higher-Order Collocates

1. discourse collocates
   - straightforward generalization with respect to textual co-occurrence
   - look at co-occurrence frequencies of tweets that were identified to be part of the discourse at hand (topic + attitude)
   - collocates represent lexical items that play a role in the **discourse**
2. second-order topic-collocates (or attitude-collocates)
   - look at co-occurrence frequencies of one set of lexical items *c* in tweets that are about a certain topic *t*
   - collocates of *c* that are particulary important for *t*

Extension: Visualization

- based on high-dimensional word embeddings (Word2Vec) (Mikolov et al., 2013)
  - basis: 133,526,833 deduplicated and preprocessed Japanese tweets collected between February 2017 and June 2018 via the Streaming API
- t-distributed stochastic neighbour-embedding (t-SNE) to project onto two-dimensional plane (van der Maaten and Hinton, 2008)
  - semantically similar items are pre-grouped together
- size of lexical items represents association strength towards (topic) node

2012.11.07 – 2012.12.24 node: 626.11 tw.p.m (5062/8084830)



Case Study: Fukushima Effect

Mass media in the aftermath of 3/11 (Heinrich et al., 2018)

German (FAZ)

- salience of *energy transition* discourse relatively stable (2011–2014)
- *nuclear phase-out* (Atomausstieg) as part of this discourse: sparked shortly after 3/11
    - political actors and issues (*Ethikkommission*, *electricity supply*)
    - economic actors (*RWE*)
    - technological issues (*Stromnetz*)

Japanese (Yomiuri)

- *nuclear phase-out* (脱原発) in 2011:
    - political actors (菅, 野田, 首相)
    - economic issues (発電, 稼働, 復興)
    - technological aspects (安全, 燃料)
- *nuclear phase-out* in 2014:
    - elections and politics (演説, as used in 街頭演説)
    - fewer words regarding economics (note アベノミクス)

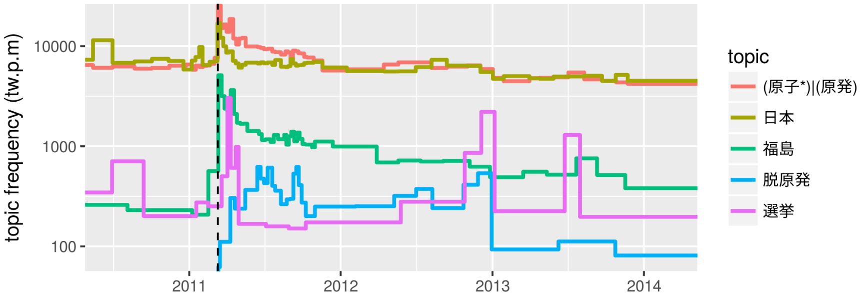


Figure: Frequencies (in tweets per million) of selected topics during the observation period on a logarithmic scale. The dashed line represents March 11, 2011. All observed frequencies peak at or shortly after 3/11.

2011.03.12 – 2011.03.19 node: 5121.9 tw.p.m (29425/5744937)

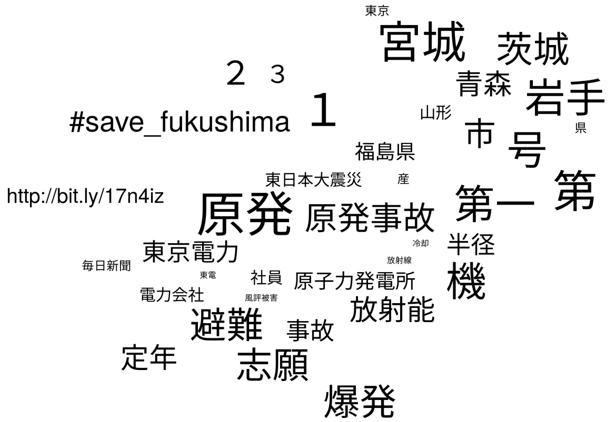


Figure: Node: 福島 (*Fukushima*).

2011.03.20 – 2011.03.26 node: 3170.69 tw.p.m (17346/5470729)

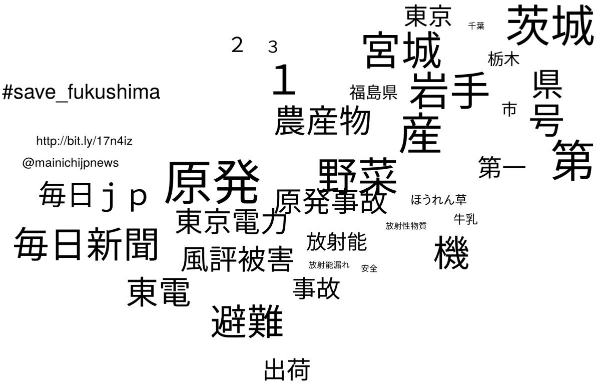


Figure: Node: 福島 (*Fukushima*).

2011.04.11 – 2011.04.18 node: 3637.94 tw.p.m (19960/5486618)

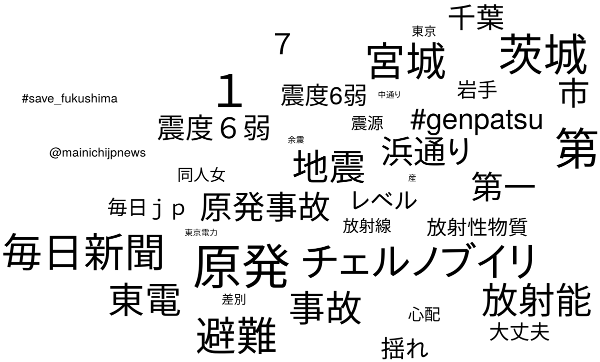


Figure: Node: 福島 (*Fukushima*).

2012.11.07 – 2012.12.24 node: 626.11 tw.p.m (5062/8084830)

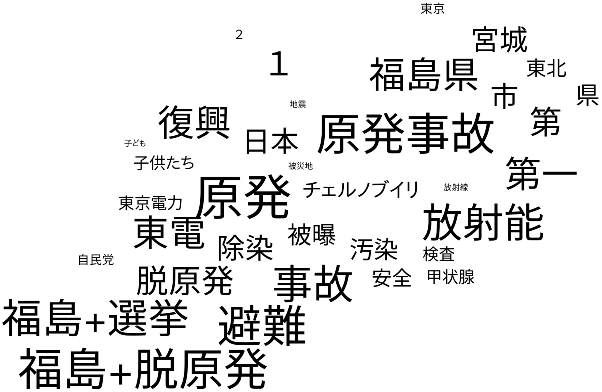


Figure: Node: 福島 (*Fukushima*).

2013.07.22 – 2013.09.09 node: 757.58 tw.p.m (5152/6800569)

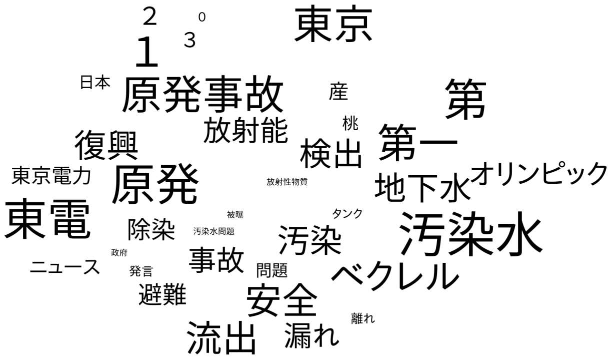


Figure: Node: 福島 (*Fukushima*).

2011.03.20 – 2011.04.02   node: 500.94 tw.p.m (5470/10919458)

Figure: Node: 選挙 (*elections*).

2011.04.03 – 2011.04.10 node: 3033.63 tw.p.m (16707/5507265)

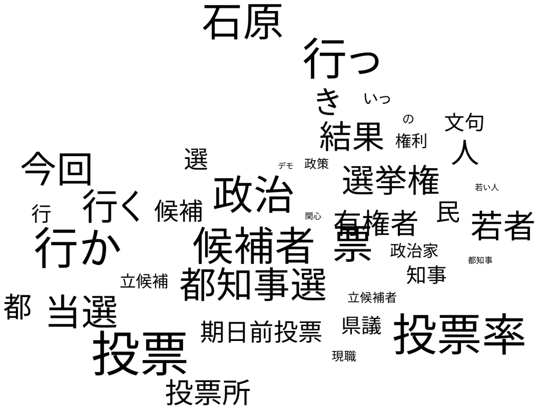


Figure: Node: 選挙 (*elections*).

2012.12.05 – 2013.01.05 node: 2208.61 tw.p.m (11047/5001799)

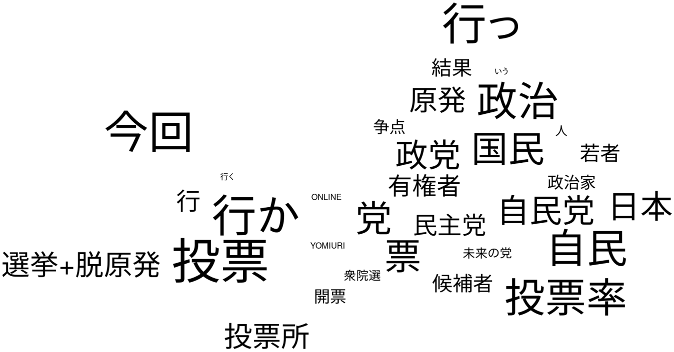
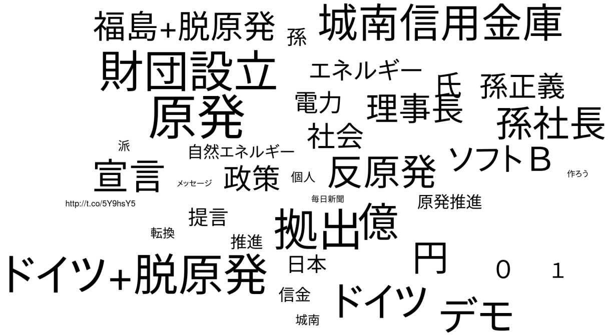


Figure: Node: 選挙 (*elections*).

2011.04.10 – 2011.04.20 node: 304.28 tw.p.m (2372/7795474)



#genpatsu

Figure: Node: 脱原発 (*phasing out nuclear energy*).

2011.06.14 – 2011.06.21 node: 622.88 tw.p.m (3613/5800493)

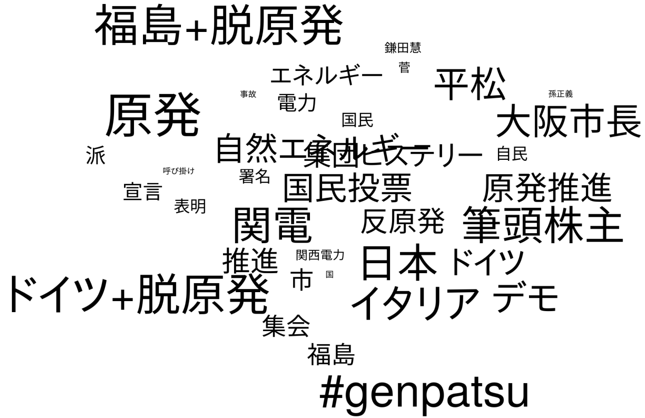


Figure: Node: 脱原発 (*phasing out nuclear energy*).

2012.11.27 – 2012.12.29 node: 536.97 tw.p.m (2858/5322440)

Figure: Node: 脱原発 (*phasing out nuclear energy*).

2009.08.14 – 2011.03.11 topic: 7568.73 tw.p.m (613835/81101453)

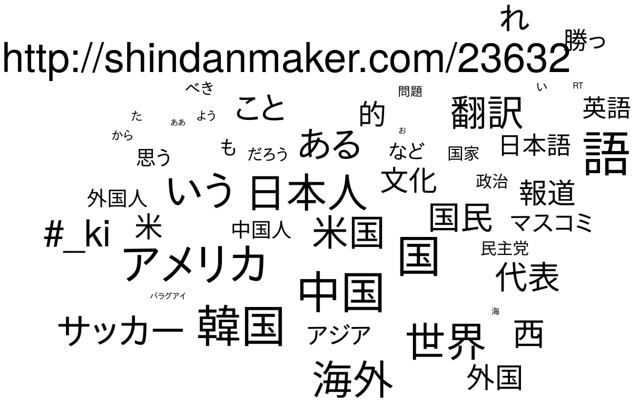


Figure: Node: 日本 (*Japan*).

2011.03.12 – 2011.12.31 topic: 7434.39 tw.p.m (1116055/150120577)

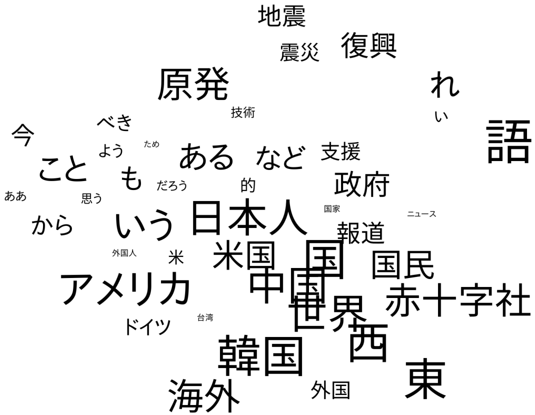


Figure: Node: 日本 (*Japan*).

2012.01.01 – 2015.03.05 topic: 5382.9 tw.p.m (502198/93294994)

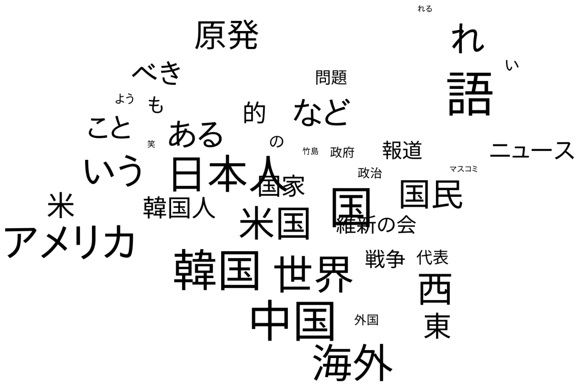


Figure: Node: 日本 (*Japan*).

2009.08.14 – 2011.03.11   topic: 113.21 tw.p.m (9096/80347136)

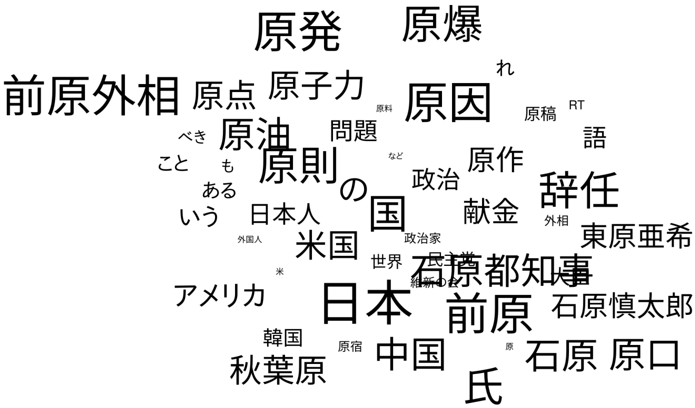


Figure: Discourse Node: 日本 (*Japan*) + (原子*)|(原発) (*nuclear energy*).

2009.08.14 – 2011.03.11 topic: 113.21 tw.p.m (9096/80347136)

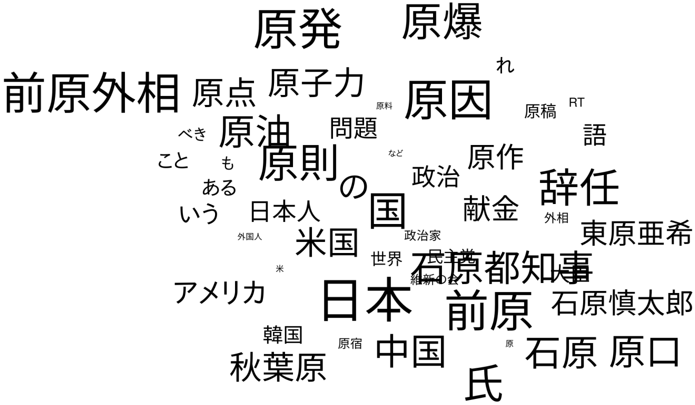


Figure: Discourse collocates of 日本 (*Japan*) + (原子*)|(原発) (*nuclear energy*).

2011.03.12 – 2011.12.31 topic: 475.92 tw.p.m (71439/150108634)

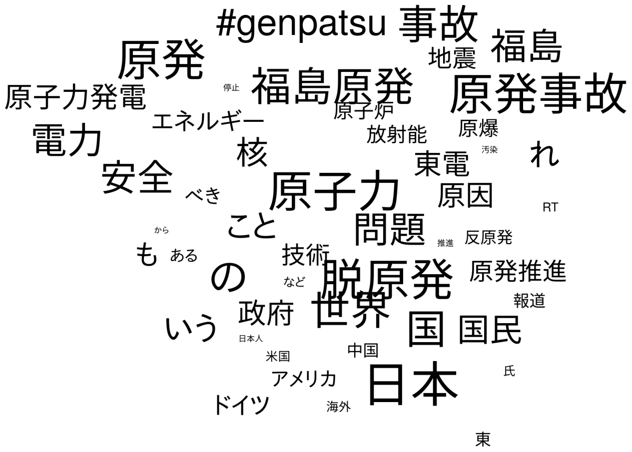


Figure: Discourse collocates of 日本 (*Japan*) + (原子*)|(原発) (*nuclear energy*).

2012.01.01 – 2015.03.05 topic: 191.53 tw.p.m (17676/92286677)

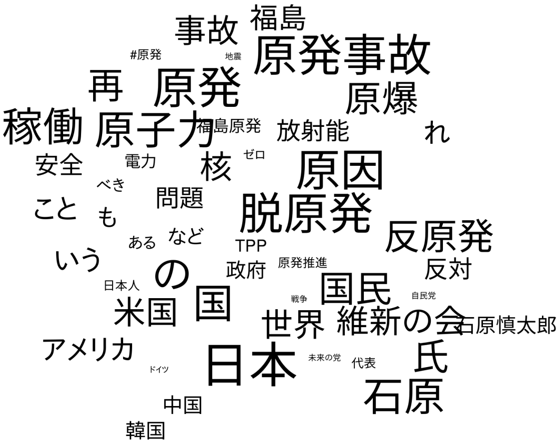


Figure: Discourse collocates of 日本 (*Japan*) + (原子*)|(原発) (*nuclear energy*).

2009.08.14 – 2011.03.11 topic: 2381.75 tw.p.m (1462/613835)

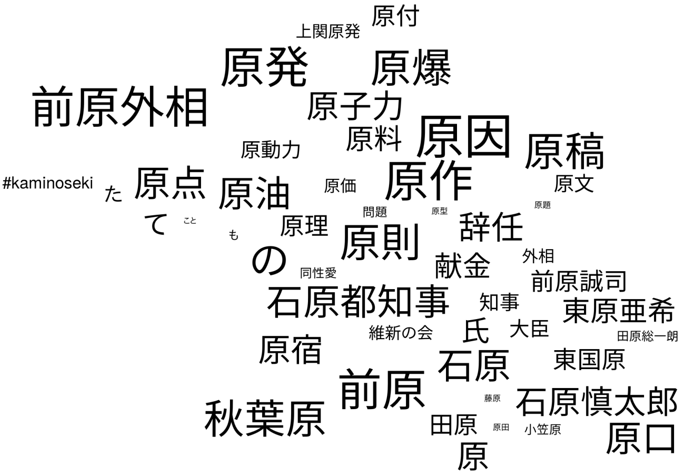


Figure: Second-order topic-collocates of 日本 (*Japan*) in tweets containing (原子*)|(原発) (*nuclear energy*).

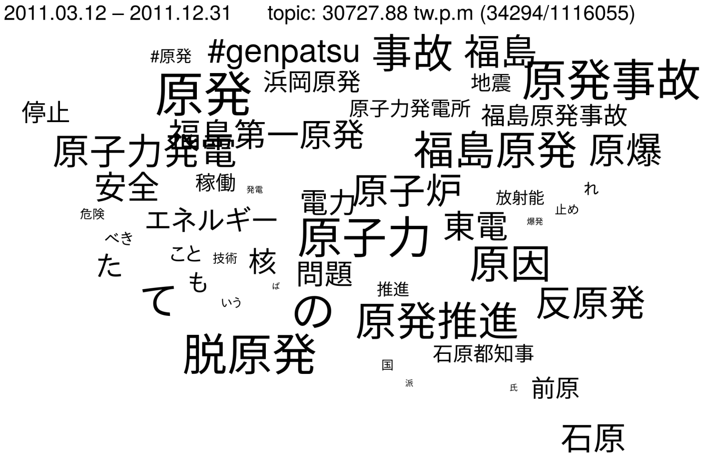


Figure: Second-order topic-collocates of 日本 (*Japan*) in tweets containing (原子*)|(原発) (*nuclear energy*).

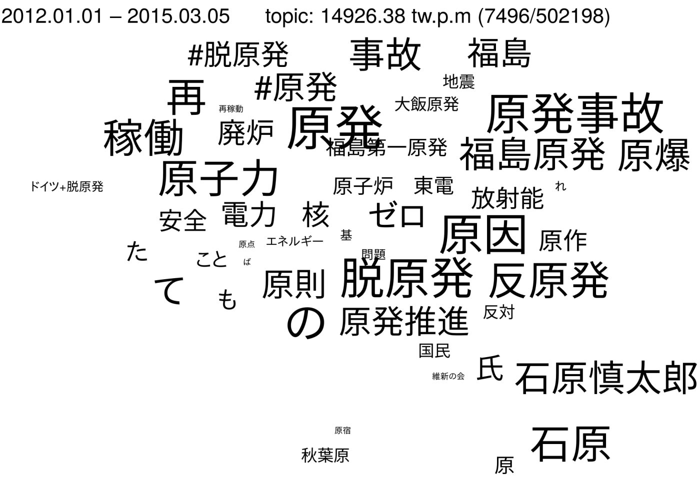


Figure: Second-order topic-collocates of 日本 (*Japan*) in tweets containing (原子*)|(原発) (*nuclear energy*).

Conclusion

Qualitative Summary

- 福島 (*Fukushima*)
  - has always been a topic on Twitter
  - important collocates during the observation 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*)
  - 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*)
- 選挙 (*elections*)
  - huge peaks in the number of tweets at dates which coincide e. g. with the elections of Tokyo's governor after resignation of 石原 (*Shintaro Ishihara*)
  - further important collocates are 結果 (*results*), 都知事選 (*gubernatorial election*), and 候補(者) (*candidate, candidacy*)
  - end of 2012: most important collocates have shifted towards 自民 (*Liberal Democratic Party*), nuclear power (plants) (原発)
  - actors change

Qualitative Summary (ctd.)

- 脱原発 (*nuclear phase-out*)
  - enters the debate only a couple of weeks after 3/11
  - whether or not to "break with nuclear energy" is a discussion led elsewhere, e. g. in ドイツ (*Germany*)
  - further important collocates are 福島 (*Fukushima*), 原発 (*nuclear power plant*), and デモ (*demonstration*)
  - another peak in the end of 2012, with political actors as collocates such as 未来の党 (the *Tomorrow Party of Japan*) and 山本太郎 (*Tarō Yamamoto*)
- 日本 (*Japan*) and (原子*)|(原発) (*nuclear energy*)
  - before 3/11: collocates of Japan mostly general (語, other countries)
  - in the aftermath of 3/11: 地震 (*earthquake*), 復興 (*reconstruction*), 原発 (*nuclear power plant*), and 赤十字社 (*red cross*)
  - after 2012: 原発 (*nuclear power plant*) remains an important collocate

Conclusion and Future Work

- CDA of Japanese Twitter data in the aftermath of 3/11
- focus on methodological advancement of the field
  - visualization (ease manual labour)
  - higher-order collocates (triangulate semantics of discourses)
- qualitative empirical level:
  - **nuclear phase-out debate** entered Japanese Twitter only **several weeks after 3/11**
  - salience of discussions about **phasing out nuclear energy** and about nuclear energy in general is quite volatile and **correlates i. a. with elections**
  - **particular parts of the nuclear energy discussion** entered the collocational profile of the very **general discourse** around Japan
- where do we go from here?

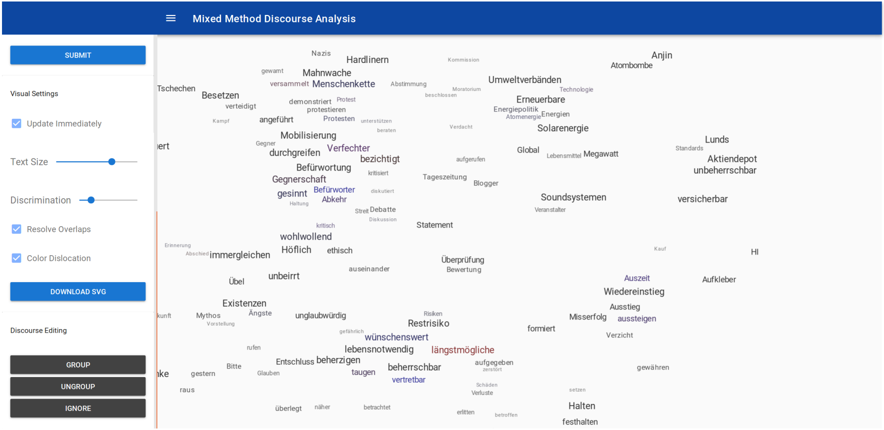
Conclusion and Future Work

- CDA of Japanese Twitter data in the aftermath of 3/11
- focus on methodological advancement of the field
  - visualization (ease manual labour)
  - higher-order collocates (triangulate semantics of discourses)
- qualitative empirical level:
  - **nuclear phase-out debate** entered Japanese Twitter only **several weeks after 3/11**
  - salience of discussions about **phasing out nuclear energy** and about nuclear energy in general is quite volatile and **correlates i. a. with elections**
  - **particular parts of the nuclear energy discussion** entered the collocational profile of the very **general discourse** around Japan
- where do we go from here?

EMERGING FIELDS INITIATIVE

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Towards Mixed-Methods Discourse Analysis



Thanks for listening.

Questions?

References

Paul Baker. *Using Corpora in Discourse Analysis*. Continuum, London, 2006.

Stefan Evert. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, chapter 58. Mouton de Gruyter, Berlin, 2008.

Michel Foucault. *L'Archéologie du savoir*. Éditions Gallimard, Paris, 1969.

Ikuo Gono'i. 2015-nen ANPO, Minshushugi wo futatabi hajimeru wakamono-tachi (ANPO in 2015. The Youth that is restarting Democracy), 2015.

Philipp Heinrich, Christoph Adrian, Olena Kalashnikova, Fabian Schäfer, and Stefan Evert. A Transnational Analysis of News and Tweets about Nuclear Phase-Out in the Aftermath of the Fukushima Incident. In Andreas Witt, Jana Diesner, and Georg Rehm, editors, *Proceedings of the LREC 2018 "Workshop on Computational Impact Detection from Text Data"*, Paris, 2018. ELRA.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

Thomas Proisl. SoMeWeTa: A Part-of-Speech Tagger for German Social Media and Web Texts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18)*, 2018.

Thomas Proisl and Peter Uhrig. SoMaJo: State-of-the-art tokenization for German web and social media texts. In Paul Cook, Stefan Evert, Roland Schäfer, and Egon Stemle, editors, *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin, 2016. Association for Computational Linguistics.

Toshinori Sato, Taiichi Hashimoto, and Manabu Okumura. 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, 2017.

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

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