Extended Abstract. In Book of Abstracts of the CLARIN Annual Conference 2019, Leipzig, Germany, 2019.

# Application of a Topic Model Visualisation Tool to a Second Language

**Maria Skeppstedt**[1,2,*], **Magnus Ahltorp**[1], **Andreas Kerren**[3], **Rafal Rzepka**[2,4], **Kenji Araki**[2]
[1]The Language Council of Sweden, the Institute for Language and Folklore, Sweden
{maria.skeppstedt,magnus.ahltorp}@sprakochfolkminnen.se
[2]Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan
{rzepka,araki}@ist.hokudai.ac.jp
[3]Department of Computer Science and Media Technology, Linnaeus University, Växjö, Sweden
andreas.kerren@lnu.se
[4]RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan

## Abstract

We explored adaptions required for applying a topic modelling tool to a language that is very different from the one for which the tool was originally developed. The tool, which enables text analysis on the output of topic modelling, was developed for English, and we here applied it on Japanese texts. As white space is not used for indicating word boundaries in Japanese, the texts had to be pre-tokenised and white space inserted to indicate a token segmentation, before the texts could be imported into the tool. The tool was also extended by the addition of word translations and phonetic readings to support users who are second-language speakers of Japanese.

## 1 Introduction and background

Topic modelling provides a means of extracting a relevant subset of documents from collections that are too large to make a fully manual analysis of all its documents feasible. The extracted documents are organised into groups by the topic modelling algorithm, each group corresponding to a topic that occurs frequently in the document collection. This ability to, in an unsupervised fashion, extract and topically sort relevant documents has been used to perform qualitative text analysis in social science and humanities research (Baumer et al., 2017).

We have previously presented a tool for visualising the output of topic modelling, which we call Topics2Themes (Skeppstedt et al., 2018a; Skeppstedt et al., 2018b). There are several tools for visualising topic modelling output, for instance with the focus on assessing and improving the quality of the topic model produced (Chuang et al., 2012; Lee et al., 2012; Jaegul Choo et al., 2013; Hoque and Carenini, 2015; Lee et al., 2017; Smith et al., 2018; Cai et al., 2018), and with the focus on supporting the user in exploring and interpreting the texts included in the document collection (Alexander et al., 2014).

Although the output of topic models in the form of a selection and sorting of documents has been shown useful for speeding up and facilitating qualitative text analysis, previous research has shown the need for users to identify other pieces of information than the automatically created topics. For instance, to identify reoccurring themes in the texts extracted, which are more thematically detailed than automatically identified topics (Baumer et al., 2017). Topics2Themes, therefore, does not only include functionality for letting the user explore and interpret the automatically extracted topics and documents, but places an equal emphasis on allowing the user to add, and subsequently explore, an additional layer of analysis. This is carried out by enabling the creation of user-defined themes that can be associated with the documents extracted by the topic modelling algorithm.

We originally created Topics2Themes for English texts. Despite the unsupervised nature of the topic modelling algorithm, which makes Topics2Themes fairly language-independent, it is not self-evident that the tool can be applied as-is to text written in a language that is typologically very different from English. To investigate this, we applied the tool to texts written in Japanese, i.e., a language that is both morphologically and orthographically different from English.



[*] International Research Fellow of Japan Society for the Promotion of Science (Postdoctoral Fellowships for Research in Japan (Short-term))

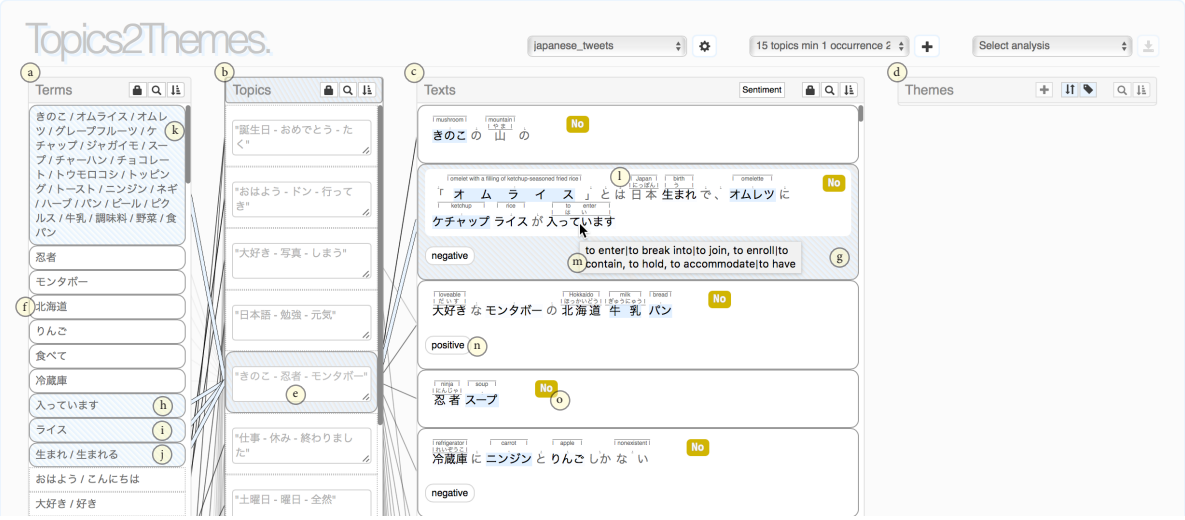


Figure 1: (a-d) The *Terms*/*Topics*/*Texts*/*Themes* panels. (e) The selected topic. (f) Example of rounded border indicating terms and texts associated with the selected topic. (g) The text over which the mouse hovers. (h-k) Terms associated with the text over which the mouse hovers. (k) Cluster of food-related words. (l) Language support in the form of phonetic reading and English translation. (m) Additional English translations for the word over which the mouse hovers. (n) A static label attached to the text by the word list matching. (o) A label that the user can change, here given an initial neutral value.

In addition, we envisioned the situation in which the text analysis of the Japanese texts would be performed by an analyst that would require some level of language support for fully understanding the texts. Such a situation would most naturally occur in a language learning situation, i.e., a situation in which the interaction with the documents is the primary reason to use the tool, and the output of the analysis is only of secondary importance. This situation could, however, also occur in the case in which a second-language speaker needs an understanding of the important content of a document collection, without having the means of employing the help of a more proficient speaker of the language. With this situation in mind, we incorporated a system into Topics2Themes that helps second-language speakers of Japanese to understand Japanese text.

## 2 Adaption to Japanese and the addition of reading support

Topics2Themes[1] uses a very simple tokenisation based on the occurrence of white space. As white space is not normally used in Japanese to indicate word boundaries, another technique for tokenisation is required. We decided not to change the tokenisation method built into Topics2Themes, but to instead require the texts imported into the tool to be pre-tokenised and white space inserted into the texts to indicate token segmentation. The tokenisation included in Topics2Themes could therefore be used as-is. For this pre-processing, we applied segmentation using the MeCab tool (Kudo, 2006), and then merged segments to tokens by matching them to the JMDict dictionary (JMdict, 2013), as implemented by Ahltorp (2012).

We also configured the tool to use Japanese stop words[2], instead of using English ones, as well as to use a word2vec model trained on a Japanese corpus to perform concept clustering. That is, Topics2Themes can be configured to apply dbscan clustering on word2vec vectors corresponding to the words in the corpus, and let all words belonging to the same cluster be collapsed into one concept, before the text is submitted to the topic modelling algorithm. For performing the clustering on Japanese, we configured Topic2Themes to use vectors from a word2vec model[3] that had been trained on Japanese texts, which

[1]The code for the Topics2Themes tool is available at: https://github.com/mariask2/topics2themes, and the code for the Japanese pre-processing can be obtained by contacting the authors.

[2]We used stopwords from: https://github.com/stopwords/japanese-stopwords/blob/master/data/japanese-stopwords.txt and extended them by frequent non-content words in the corpus used.

[3]https://github.com/shiroyagicorp/japanese-word2vec-model-builder

had been segmented by MeCab and merged with the help of a dictionary.

For reading support, we incorporated a system constructed for Japanese language learning that provides English translations for the tokens included in the text as well as phonetic readings (*furigana*) for the *kanji*[4] characters (Ahltorp, 2012). Topics2Themes was extended to use the *ruby*-tag provided in HTML to display the phonetic reading and one English translation in a small font above each token. In addition, when the user hovers the mouse over a token, all available English translations are shown in the form of a tooltip. To further help the reader, we matched the texts to Japanese sentiment and emotion word lists (Nakamura, 1993; Takamura et al., 2005; Rzepka and Araki, 2012; Rzepka and Araki, 2017), to be able to indicate the existence of such words in the text.

## 3 Application of the adapted tool on a Japanese corpus

We applied the extended version of Topics2Themes on a corpus consisting of 1,000 microblogs[5] collected with the criterium that they should contain the same content written in Japanese and in English (Ling et al., 2014). The tool was applied on the Japanese part of the microblogs.

We configured Topics2Themes to try to find 15 topics among the 1,000 texts and to run the topic modelling 100 times, only keeping topics that occurred in all re-runs. This resulted in 12 stable topics being identified by the tool. The most prominent among those can be seen in the *Topics* panel in Figure 1, where each topic is represented by its three most closely associated terms. The small corpus size used, and the small size of each text in the corpus, might make it difficult for the topic modelling algorithm to find reoccurring topics. We therefore configured the tool to allow a large maximum distance for the word2vec-based concept clustering, i.e., two words with a Minkowski distance of up to 0.7 could be included in the same cluster. This makes it possible for the topic model algorithm to find topics based on semantically related words, e.g., on the cluster of food-related words shown in the top element of the *Terms* panel in Figure 1.

Figure 1 also indicates how the results can be explored by the Topics2Themes tool. In the situation shown in the figure, the user has double-clicked on, and thereby selected, the fifth topic in the *Topics* panel. This has had the effect that the terms most closely associated with the selected topic have been sorted as the top-ranked elements in the *Terms* panel, and that the texts most closely associated with the topic have been sorted as the top-ranked elements in the *Texts* panel. The elements associated with the selected topic have also been given a bold, rounded border. The figure also shows how the user hovers the mouse over one of the texts, which has the effect that the terms included in this text, as well as the topic(s) to which the text is associated, are highlighted with a blue colour.

The language support, in the form of phonetic reading and English translation, is shown in a small font above the Japanese texts, as well as in the form of a tool tip for the word over which the mouse hovers. The *Texts* panel also displays the output of the sentiment and emotion word list matching, in the form of labels attached to the texts.

By inspecting the top-ranked terms and texts for each topic, it can be concluded that reoccurring themes in this corpus include greetings, language studies, events in Japan, food, and natural disasters. To further explore recurring themes, the texts selected by the algorithm as most closely associated with the 12 extracted topics should be manually analysed. Such an analysis, using the *Themes* panel for documentation of themes identified, will be included in future work. We intend to let a learner of Japanese perform the analysis, in order to also obtain an indication of the usefulness of the language support provided in our extension of the Topics2Themes tool.

## 4 Acknowledgements

This research was funded by the Japan Society for the Promotion of Science, and will continue within the Språkbanken and SWE-CLARIN infrastructures, supported by the Swedish Research Council (2017-00626).

[4]The logographic Chinese characters adapted to and used in Japanese.

[5]The corpus used is listed as a CLARIN resource at: https://www.clarin.eu/resource-families/parallel-corpora, and is also available at: http://www.cs.cmu.edu/~lingwang/microtopia/#twittergold

## References


Magnus Ahltorp. 2012. A Personalizable Reading Aid for Second Language Learners of Japanese. Master's thesis, Royal Institute of Technology.

Eric Alexander, Joe Kohlmann, Robin Valenza, Michael Witmore, and Michael Gleicher. 2014. Serendip: Topic model-driven visual exploration of text corpora. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 173–182, Oct.

Eric P. S. Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K. Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, 68(6):1397–1410, June.

Guoray Cai, Feng Sun, and Yongzhong Sha. 2018. Interactive visualization for topic model curation. In *Joint Proceedings of the ACM IUI 2018 Workshops*. CEUR-WS.

Jason Chuang, Christopher D. Manning, and Jeffrey Heer. 2012. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 74–77, New York, NY, USA. ACM.

Enamul Hoque and Giuseppe Carenini. 2015. Convisit: Interactive topic modeling for exploring asynchronous online conversations. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, pages 169–180, New York, NY, USA. ACM.

Chandan K. Jaegul Choo, Haesun Changhyun Lee, Haesun Reddy, and Haesun Park. 2013. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):1992–2001.

JMdict. 2013. The JMDict Project. http://www.edrdg.org/jmdict/j_jmdict.html.

Taku Kudo. 2006. Mecab : Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.jp*.

Hanseung Lee, Jaeyeon Kihm, Jaegul Choo, John Stasko, and Haesun Park. 2012. iVisClustering: An interactive visual document clustering via topic modeling. *Computer Graphics Forum*, 31(3pt3):1155–1164.

Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. 2017. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*.

Wang Ling, Luis Marujo, Chris Dyer, Alan Black, and Isabel Trancoso. 2014. Crowdsourcing high-quality parallel data extraction from twitter. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, WMT '14, Stroudsburg, PA, USA. Association for Computational Linguistics.

Akira Nakamura. 1993. *Kanjo hyogen jiten [Dictionary of Emotive Expressions]*. Tokyodo Publishing.

Rafal Rzepka and Kenji Araki. 2012. Polarization of consequence expressions for an automatic ethical judgment based on moral stages theory. *IPSJ SIG Notes*, 14(2012-NL-207):1–4.

Rafal Rzepka and Kenji Araki. 2017. What people say? web-based casuistry for artificial morality experiments. In *International Conference on Artificial General Intelligence*, pages 178–187. Springer.

Maria Skeppstedt, Andreas Kerren, and Manfred Stede. 2018a. Vaccine hesitancy in discussion forums: Computer-assisted argument mining with topic models. In *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth*, number 247 in Studies in Health Technology and Informatics, pages 366–370. IOS Press.

Maria Skeppstedt, Kostiantyn Kucher, Manfred Stede, and Andreas Kerren. 2018b. Topics2Themes: Computer-Assisted Argument Extraction by Visual Analysis of Important Topics. In *Proceedings of the LREC Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources*, pages 9–16.

Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Intelligent user interfaces. In *User-Centered Design and Evaluation of a Human-in-the-Loop Topic Modeling System*.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 133—140.