Topic modelling applied to a second language: A language adaptation and tool evaluation study

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

Abstract

The Topics2Themes tool, which enables text analysis on the output of topic modelling, was originally developed for the English language. In this study, we explored and evaluated adaptations required for applying the tool to Japanese texts. That is, we adapted Topics2Themes to a language that is very different from the one for which the tool was originally developed. To apply Topics2Themes to Japanese texts, in which white space is not used for indicating word boundaries, the texts had to be pre-tokenised and white space inserted to indicate a token segmentation. Topics2Themes was also extended by the addition of word translations and phonetic readings to support users who are second-language speakers of Japanese. To evaluate the adaptation to a second language, as well as the reading support, we applied the tool to a corpus consisting of short Japanese texts. Twelve different topics were automatically identified, and a total of 183 texts representative for the twelve topics were extracted. A learner of Japanese carried out a manual analysis of these representative texts, and identified 35 reoccurring, fine-grained themes.

1 Introduction and background

Topic modelling provides a means of extracting a relevant subset of texts from a document collection that is too large to make a fully manual analysis of all its texts feasible. The extracted texts are organised into groups by the topic modelling algorithm, each group corresponding to an automatically detected topic that occurs frequently in the document collection (Blei et al., 2003; Blei, 2012; Jelodar et al., 2019). In addition to being associated to a group of extracted texts, the topics detected are also represented by a list of terms that are associated with the topics. This ability to extract and topically sort relevant texts in an unsupervised fashion has been used to perform qualitative text analysis in social science and humanities research (Baumer et al., 2017).

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; Choo et al., 2013; Hoque and Carenini, 2015; Lee et al., 2017; Cai et al., 2018; Smith 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). A popular topic modelling visualisation approach is, for example, to display the topics and their associated texts or terms in a grid, and to use visual markers such as circles of different sizes and colours to indicate the level of association between a topic and a text or term (Chuang et al., 2012; Alexander et al., 2014).

Maria Skeppstedt, Magnus Ahltorp, Kostiantyn Kucher, Andreas Kerren, Rafal Rzepka and Kenji Araki 2020. Topic modelling applied to a second language: A language adaptation and tool evaluation study. *Selected papers from the CLARIN Annual Conference 2019*. Linköping Electronic Conference Proceedings 172: 172 145–156.

The output of topic models, in the form of an automatic selection of subsets of texts and terms from a large text collection, has been shown useful for speeding up and facilitating qualitative text analysis (Baumer et al., 2017). Previous research has, however, also demonstrated that relying only on extracted terms—without also analysing extracted texts—has lead to misunderstandings regarding the content of the text collection (Lee et al., 2017). In addition, there is not always a one-to-one correspondence between (i) the topics automatically extracted by the topic modelling algorithm, and (ii) what the user identifies as interesting, reoccurring categories of information when analysing a text collection (Baumer et al., 2017). Baumer et al. compared two methods for extracting reoccurring information in 2,190 free-text survey responses: (i) the use of topic modelling for selecting reoccurring topics, and (ii) a fully manual approach, in which a grounded theory-based analysis was applied. For the manual approach, all survey response texts were analysed in the search for what the authors call *themes*, i.e., categories formed by reoccurring information found in the texts. When comparing the output from the topic modelling and the grounded theory-based analysis, it could be concluded that the "topic modeling results captured to a surprising degree many of the themes identified in grounded theory, and vice versa." However, topics produced by the topic modelling algorithm often corresponded to several of the themes detected in the manual analysis, and some of the manually detected themes could be associated with several topic modelling-produced topics. With the aim of helping the user deal with this possible difference in granularity between automatically detected topics and manually detected themes, we have previously developed the Topics2Themes visualisation tool. The tool facilitates a manual search for themes, among the texts selected by the topic modelling algorithm (Skeppstedt et al., 2018a; Skeppstedt et al., 2018b).

With the Topics2Themes tool, we have thereby expanded the functionality typically provided by previous tools. The user can not only explore and interpret the automatically extracted topics and texts, but also 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 texts extracted by the topic modelling algorithm. These user-defined themes and their text and topic associations, as well as their associations to automatically extracted terms, can then be explored in the tool. Thereby, an overview of the text analysis can be obtained, in which the automatically extracted information is integrated with the output of the manual analysis performed by the user. We also provide functionality for including metadata in the topic model visualisation in the form of text labels. The labels are either static or take the form of dynamic text labels that can be changed by the user.

We originally created Topics2Themes for English texts. Despite the unsupervised nature of the topic modelling algorithm, which makes the functionality of 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.

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 texts 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 the situation of a language learner in mind, we incorporated a system into Topics2Themes that helps second-language speakers of Japanese to understand Japanese text. We are not aware of any previous tools that combine the possibilities of using topic modelling for extracting and sorting the most relevant information from large text collections, with the functionality of providing reading support for language learners.

We here describe (i) the adaptation of the Topics2Themes tool to Japanese and to the situation in which the tool would be used by a second-language learner of Japanese, and (ii) the evaluation of the adapted tool on a Japanese document collection. The study has resulted in a new, language-independent version of the Topics2Themes tool, in which reading support can be provided. General usability issues, detected when the tool was evaluated on Japanese texts, were also corrected in the new version of the tool.

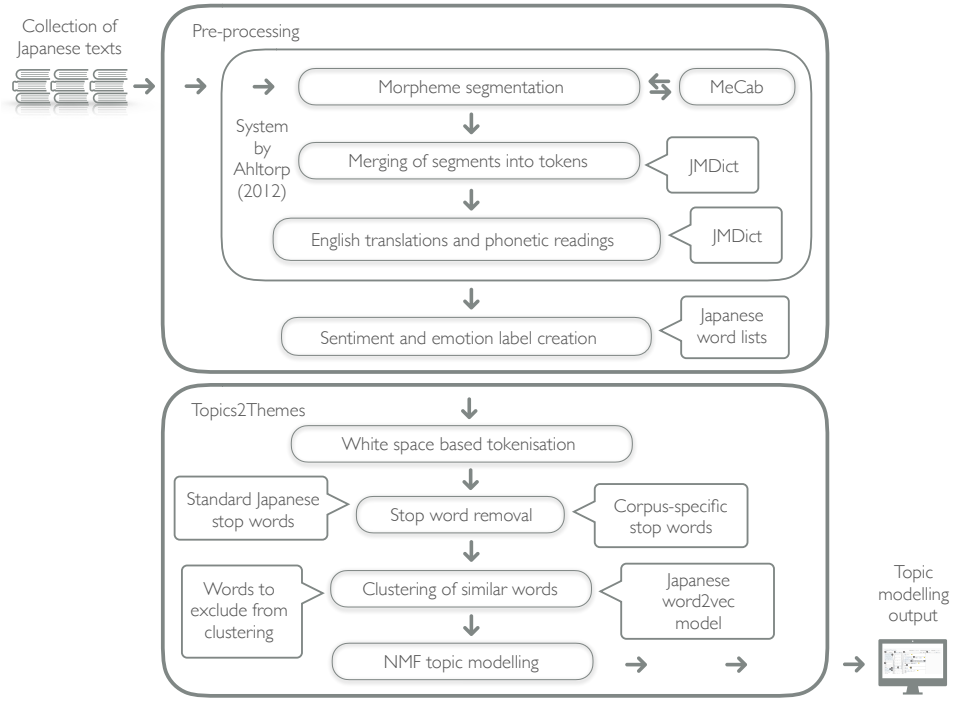

Figure 1: The components of the pre-processing and of the Topics2Themes tool adapted to Japanese. The language-specific parts consist of the entire pre-processing functionality, of the three word lists used (stop word lists, and words to exclude from clustering), and of the word2vec model used by Topics2Themes.

The development of Topics2Themes[1] was initiated by research funding from the Swedish Research Council, and the adaptation to Japanese and the evaluation on Japanese texts was funded by the Japan Society for the Promotion of Science. The updated version of the tool, in which usability issues were corrected, was developed within the Språkbanken and SWE-CLARIN infrastructures. Further developments of Topics2Themes will also be carried out within the Språkbanken and SWE-CLARIN infrastructures.

2 Method

The adaptation to Japanese consisted of adding an additional step in the process of using Topics2Themes, in the form of a pre-tokenisation of the texts before they were imported into the tool. Topics2Themes was also configured to use Japanese stop words and a word2vec (Mikolov et al., 2013) model trained on Japanese in the topic modelling process, as well as Japanese word lists for adding automatic labelling of the texts. Finally, a system for reading support for second-language speakers was incorporated.

We thereafter applied the adapted tool to a Japanese corpus, and the texts extracted by the topic modelling algorithm were then manually analysed by a learner of Japanese.

2.1 Adaptation to Japanese and the addition of reading support

Topics2Themes 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 tokenisation technique is needed. 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.

[1]The code for the Topics2Themes tool is available at: `https://github.com/mariask2/topics2themes`, and the code for the Japanese pre-processing at: `https://github.com/mariask2/T2T_pre-processing_ja`.

For the pre-tokenisation, we segmented the text into morphemes using the MeCab tool (Kudo, 2006), and then merged morphemes into tokens by matching them to the JMDict dictionary (JMdict, 2013), as implemented by Ahltorp (2012).

The Topics2Themes tool can be configured to apply DBSCAN (Ester et al., 1996) clustering on word2vec vectors that correspond to the words in the corpus. Words belonging to the same cluster can thereby be collapsed into one concept, before the text is submitted to the topic modelling algorithm. The maximum distance between two words for them to be counted as the same concept can be adjusted by the user. That is, a large maximum distance allows for not only synonyms and different morphological instantiations of the same concept to be clustered together, but also creates groups in the form of semantically related concepts. To be able to perform the clustering on Japanese, we configured Topics2Themes to use vectors from a word2vec model[2] that had been trained on Japanese texts. The texts had been segmented by MeCab, and the segments had been merged into tokens with the help of a dictionary. Further on, a list of 111 words to exclude from the automatic clustering was manually created, since the clustering grouped these words together with semantically distant ones.

We also configured the tool to use Japanese stop words. Firstly, we used a Japanese stop words list available online[3]. This list was then extended by adding 150 frequent Japanese non-content words that occurred in the corpus to which the tool was applied.

For reading support, we incorporated a system constructed for Japanese language learning that provides a ranked list of English translations for each token included in the text, as well as a phonetic reading (*furigana*) for each Japanese *kanji*[4] character in the text. This tool has been developed for, and evaluated on, beginner learners of Japanese as well as learners on an intermediate level[5] (Ahltorp, 2012).

Topics2Themes was extended to use the *ruby*-tag provided in HTML to display the phonetic reading and the top-ranked 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. The functionality provided is not specific to Japanese. Instead, the extended version of Topics2Themes will provide this kind of reading support to any input text that indicates translations and/or phonetic readings using the same HTML-format.

In an attempt to further help the reader to understand and analyse the texts, we also created metadata labels by matching the texts to Japanese sentiment and emotion word lists (Nakamura, 1993; Takamura et al., 2005; Rzepka and Araki, 2012; Rzepka and Araki, 2017). Texts that contained words present in the lists were given static labels to indicate in which list they were present, and the sentiment and emotion words present in the text were also marked with a green or red background, for positive or negative words, respectively.

Figure 1 gives an overview of the components of the pre-processing and of the resources required to run Topics2Themes on the Japanese text collection.

2.2 Application of the adapted tool to a Japanese corpus

We applied the extended version of Topics2Themes to a corpus consisting of around 1,000 microblogs[6] 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 to the Japanese part of the microblogs, and the English part of the texts was not used in this study.

We configured Topics2Themes to try to find 15 topics using the NMF (non-negative matrix factorisation) topic modelling algorithm (Lee and Seung, 2001). The tool was further configured to run the topic modelling algorithm 100 times on the text collection, and to only keep topics that were stable enough to occur in all re-runs. This resulted in 12 stable topics being identified. The most prominent among those

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

[3] https://github.com/stopwords/japanese-stopwords/blob/master/data/japanese-stopwords.txt

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

[5] More specifically, the levels A1–B1 according to the Council of Europe CEFR levels.

[6] 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.

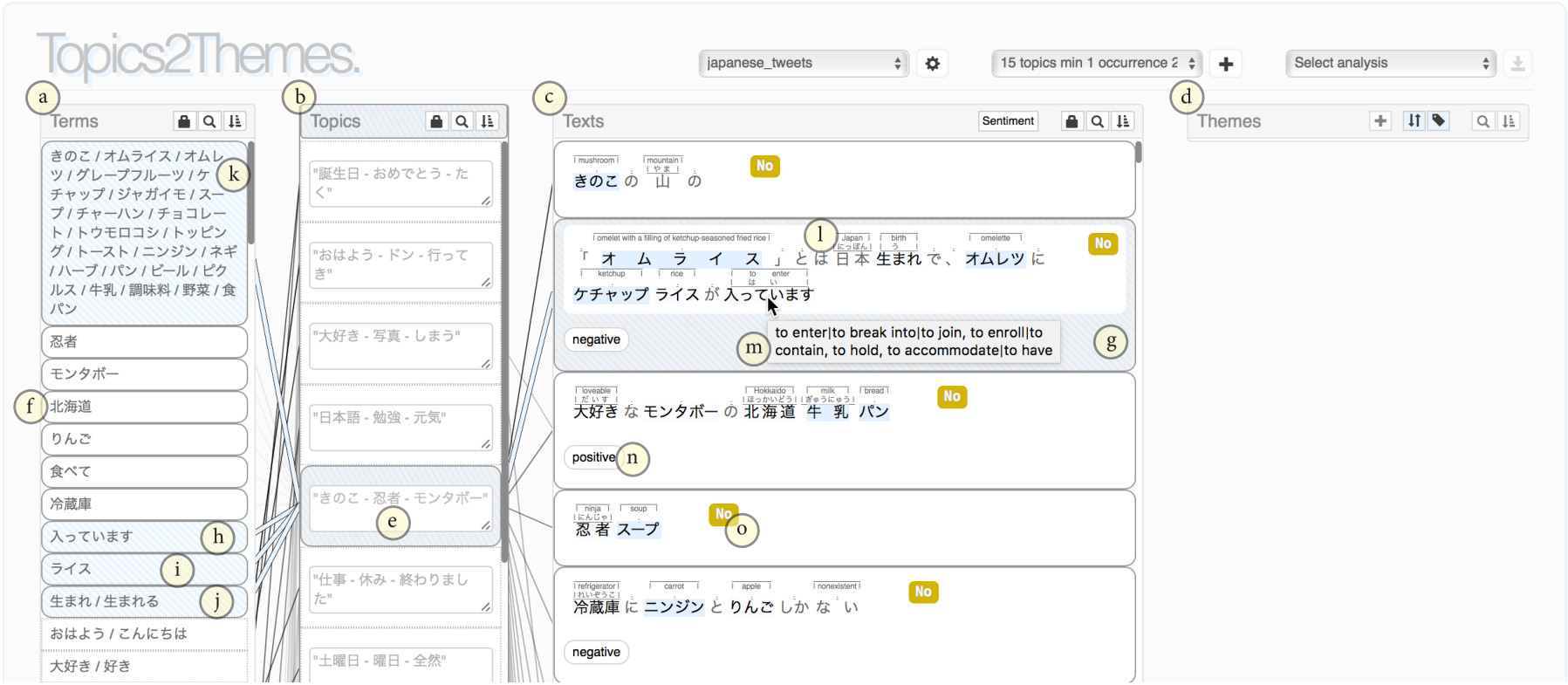

Figure 2: User interface of Topics2Themes at the early stages of analysis. (a–d) The *Terms*/*Topics*/*Texts*/*Themes* panels. (e) The selected topic. (f) Example of rounded border indicating terms and texts associated with the selected topic. (g) The text over which the mouse hovers. (h–k) Terms associated 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) Static label from automatic word list matching. (o) Dynamic label that can be changed by the user, here given an initial neutral value.

Figure 3: User interface of Topics2Themes at the later stages of analysis. (a–d) The *Terms*/*Topics*/*Texts*/*Themes* panels. (e) Example of a topic description written by the user. (f) The user has selected one topic through double-clicking on this element. (g–h) Lines showing associations between terms and topics—as well as between topics and texts—created by the topic modelling algorithm. (i) Lines showing associations between texts and themes, assigned by the user. (j) Labels indicating which themes the text has been assigned to. (k–l) Two themes that the user has assigned to texts associated with the selected topic. (m) The number of dynamic labels indicates the number of texts in which the theme occurs. (n) Static labels associated with texts in which the theme occurs. (o) One of the assigned texts contains a *positive* evaluation. (p–q) The sentiment button is selected, which results in the output of automatic sentiment word list matchings being shown. (q–r) The red sentiment markings and the static labels indicate negativity of the selected topic. (s) Button for creating a new theme.

can be seen in the *Topics* panel in Figure 2, where each topic is represented by its three most closely associated terms.

The relatively small size of the corpus 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[7] for the word2vec-based concept clustering. This makes it possible for the topic model algorithm to find topics based on semantically related words. One example of such a cluster of semantically related words is the cluster of food-related words shown in the top element of the *Terms* panel in Figure 2, which e.g., includes "mushroom", "grapefruit", "soup", "toast", "carrot", "bread", "beer", "milk" and "vegetables". Another example is given by the cluster of words for pain and diseases, which is shown in the top element in the *Terms* panel in Figure 3.

Figure 2 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 further 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 tooltip 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.

2.3 Manual analysis of the texts extracted

Topics2Themes automatically extracted a total of 183 texts from the text collection as typical for the twelve topics identified. These texts were manually analysed by a learner of Japanese. The learner of Japanese (one of the authors) had very limited experience in reading authentic Japanese texts and had previously mainly read texts from beginner-level textbooks. The goal of the analysis was to find examples of themes that reoccur in these types of Japanese-English bilingual microblogs.

The texts extracted were, for each topic, analysed with the help of the language support provided. The analysis was first carried out individually by the language learner. The same texts were, thereafter, read by the language learner with the help of a Japanese language teacher. The teacher was a native speaker of Japanese and had no previous experience in using tools similar to Topics2Themes. The content of the texts and the identified themes were discussed, and misunderstandings that had led to incorrectly identified or overlooked themes were corrected.

In addition to analysing the texts for reoccurring themes, the language learner also used the dynamic labelling functionality included in Topics2Themes for attaching the sentiment labels *positive* or *negative* to texts whose content included a positive or negative evaluation.

While using the tool, reflections on its usefulness were made, and notes regarding usability issues were taken.

3 Results and reflections

Results of the study consist of examples of themes that are reoccurring in the corpus, as well as of reflections on the usefulness and usability of the tool and its language support, when applied to Japanese texts.

3.1 Outcome of the manual analysis

Table 1 shows the outcome of the analysis task given, i.e., the task of finding examples of reoccurring themes in the collection of microblogs. The table shows the final analysis, after the Japanese teacher had corrected the analysis carried out individually by the learner. This analysis resulted in that a total of 78

[7] Two words with a Minkowski distance of up to 0.7 could be included in the same cluster.

Topic description	Theme descriptions	Nr of occ.
Birthdays	• **Birthday greetings**	12 (12)
Good morning greetings	• ***Good morning greetings***	*12 (12)*
	• *Reports on/confirmations regarding weekdays*	*1 (5)*
	• *Reports of going to work/work starting*	*1 (3)*
Expression of liking	• **Expression of liking towards a person**	6 (6)
	• **Images/photos that someone likes**	4 (4)
	• *Food that someone likes*	*1 (3)*
	• *Wise sayings and advice on how to live*	*1 (8)*
	• *Injury, illness or pain*	*1 (8)*
Studies of the Japanese and English languages, studies in general and matters related to language	• **Questions about Japanese studies**	2 (2)
	• **Doubts regarding English studies**	2 (2)
	• **Someone reports to study Japanese**	2 (2)
	• **Someone's level of English**	2 (2)
	• **Changes of texts into Japanese**	2 (2)
Food and food metaphors	• ***Food in general***	*7 (9)*
	• **Cooking and food ingredients**	4 (4)
	• ***Food that someone likes***	*3 (3)*
	• ***Food metaphors***	*3 (3)*
	• *Wise sayings and advice on how to live*	*1 (8)*
Work	• **Rest from work/reports of work that ends**	5 (5)
	• ***Reports of going to work/work starting***	*3 (3)*
	• *Good morning greetings*	*1 (12)*
Feelings and days of the week	• ***Reports on feelings***	*5 (7)*
	• ***Reports on/confirmations regarding weekdays***	*5 (5)*
	• *Good morning greetings*	*1 (12)*
	• *Information about events*	*1 (7)*
Okay	• ***Worries of whether something/oneone is okay***	*7 (7)*
	• ***Natural disasters and bad weather***	*2 (10)*
Good things and good people	• ***Wise sayings and advice on how to live***	*5 (8)*
	• **Questions on appearances/methods**	4 (4)
	• ***Reports on feelings.***	*2 (7)*
	• *Food in general*	*1 (9)*
	• *Natural disasters and bad weather*	*1 (10)*
	• *Worries of whether something/someone is okay*	*1 (7)*
	• *Food metaphors*	*1 (3)*
Information about events taking place in different cities	• ***Information about events taking place in different Japanese cities***	*6 (7)*
Injury, illness or pain	• ***Injury, illness or pain***	*7 (8)*
	• *Natural disasters and bad weather*	*1 (10)*
Natural disasters, relations with Korea and goodbye greetings	• ***Natural disasters and bad weather***	*8 (10)*
	• **Korea-Japan relations**	3 (3)
	• ***Worries of whether something is okay***	*2 (7)*
	• *Wise sayings and advice on how to live*	*1 (8)*

Table 1: Themes found in texts associated with each one of the automatically detected topics. *Nr of occ.* indicates the number of texts, associated with this topic, in which the theme was found. The number shown in parenthesis indicates the total number of texts in which this theme occurred. Bold indicates that a theme has occurred at least twice in texts associated with the topic. Italics indicates that a theme has also been found in texts associated with other topics, and thereby is listed multiple times in the table. (Note that the same text can be associated to several topics and assigned to several themes.)

themes were identified, of which 35 occurred at least twice, and 19 at least three times among the texts analysed.

The table includes all the themes that occurred at least three times, as well as the themes that occurred at least twice for the fourth topic (for which no themes occurring more than twice were identified). Themes that were assigned to texts associated with several topics are listed multiple times in the table, once for each topic. When only taking themes that occurred at least twice for a topic into account, it can be seen that most of the twelve topics extracted contained semantically coherent themes. Examples include the themes related to (i) birthday greetings, (ii) good morning greetings, (iii) food, (iv) language, (v) work starting and ending, and (vi) injury, illness and pain. There were, however, also topics with non-coherent themes, e.g., the topic "Feelings and days of the week".

Most texts did not contain a positive or negative evaluation, and hence were not assigned a positive or negative label with the manual labelling functionality provided by the dynamic labels. Exceptions were texts belonging to the themes "Images/photos that someone likes" (4 positive), "Expression of liking towards a person" (6 positive), "Food that someone likes" (4 positive), and "Food in general" (1 positive).

From the re-analysis performed together with the teacher, it could be concluded that a total of 25 texts among the 183 texts analysed had been misunderstood by the learner of Japanese. For these texts, themes assigned were removed or new theme assignments were created.

Figure 3 gives an example of what the interface of the Topics2Themes tool looks like when themes have been added. The figure shows texts associated with the topic "Injury, illness or pain" that have been assigned to four different themes created in the *Themes* panel. The figure also shows that the topics have been given user-defined descriptions, which have replaced the initial default names.

3.2 Reflections on usefulness

The subjective reflection of the analysis led to the insight that the application of Topics2Themes to the text collection enabled users to access text content that otherwise would have been very difficult to access for someone not used to reading authentic Japanese texts. Although some level of Japanese language skills were still required to access the text content, the reading support provided made it possible to focus on the search for reoccurring themes while reading the texts. That is, it was feasible to focus on the content of the texts, instead of having to use effort for manually tokenising it, for figuring out the dictionary form of the tokens, and for looking them up in a dictionary.

The automatic selection of a subset of the texts for manual analysis was also perceived as an indispensable feature for accessing the content of the document collection. It would have taken a very long time for the learner of Japanese to manually analyse all posts in the collection, in order to find examples of reoccurring themes, even with the help of the language support provided. To manually read 183 short texts with automatic reading support was, however, perceived as a feasible task for the language learner.

These subjective reflections were supported by the more objective facts that (i) only around 14 percent of the texts analysed had been misunderstood, despite the language learner's previous unfamiliarity with reading authentic Japanese texts, and (ii) by the detection of the 35 examples of reoccurring themes among the subset of 183 texts selected by the tool.

The output of the word list matching, in the form of texts being given static labels as well as in the form of polarity words in the text being highlighted in green or red, was useful for gaining a first impression of the topic. For instance, the static labels and the red highlighting of words signifying negative polarity, which are shown for the texts associated with the topic "Injury, illness or pain" in Figure 3, indicate that this topic includes negative content. The output of the word list matching was, however, not perceived as useful for performing the text analysis of each individual text.

3.3 Usability issues

Usability issues and missing features detected while performing the analysis were mostly related to the interaction with the interface of the Topics2Themes tool, as well as to how information was presented and sorted in the tool. The analysis of the Japanese texts performed in this study is one of the first authentic use cases to which the Topics2Themes tool has been applied, and we therefore expected that

usability issues not stemming from the adaptations performed for Japanese would be detected. Three large usability issues were detected.

Previously created themes relevant for the topic and texts that are being analysed are typically sorted as the top-ranked elements in the *Themes* panel. However, when the user creates a new theme element, this new element was positioned at the bottom of the *Themes* panel in the version of Topics2Themes used for the evaluation. This causes unnecessary scrolling when the newly created theme is to be used, and the tool's functionality was therefore changed to position newly created themes as the first element in the *Themes* panel.

The content of the *Terms* panel was used in the process of determining (i) which additional words should be added to the stop word list, (ii) which words to add to the list of words to be excluded from the automatic clustering process. This panel was also used for gaining an initial overview of the output of the topic modelling algorithm. However, the *Terms* panel was not perceived as useful when performing the actual analysis of the texts. At the same time, the horizontal space available for writing theme descriptions was perceived as too small. Functionality for minimising the *Terms* panel was therefore added, to make it possible for the user to choose to have more screen space available for the *Themes* panel while performing the text analysis.

The evaluated version of the tool did not include any functionality for determining whether there were texts in the *Texts* panel, for which no theme assignments had yet been made by the user. Omitted texts were therefore difficult to detect. A functionality for sorting texts according to their number of assigned themes was therefore added to the *Texts* panel, i.e., a functionality that can be used for easily finding texts without any theme associations.

There were also a number of smaller usability issues associated with the lists being automatically resorted or being automatically scrolled to the top element in cases when these events should not take place. These issues have all been corrected, and a new version of Topics2Themes has been released with the implemented corrections.

4 Conclusions

The aim of the Topics2Themes tool is to provide functionality for automatic extraction and sorting of a subset of texts from a text collection too large for a fully manual analysis. The automatically extracted texts are to contain examples of themes that reoccur in the text collection, and Topics2Themes also provides functionality for documenting reoccurring themes detected when these texts are manually analysed. The main goal of the current study was to investigate whether it is possible to achieve this aim when Topics2Themes is applied to a language very different from English, i.e., different from the language for which the tool was originally developed. When applying the tool to a collection of around 1,000 short Japanese texts, and manually analysing 183 of these texts, 35 examples of reoccurring themes were identified. 19 of these themes occurred at least three times among the extracted texts. This shows that the functionality of extracting relevant texts can be achieved also when Topics2Themes is applied to texts written in another language.

The current study also included an evaluation of the usefulness of the Japanese extension of Topics2Themes for making it possible for a learner of Japanese to access the content of a collection of authentic Japanese texts. The learner perceived the Japanese reading support provided by the tool, and the fact that only a subset of a large text collection was extracted for manual analysis, as necessary for accessing the content of the text collection. These subjective reflections were supported by the fact that only around 14 percent of the texts analysed had been misunderstood by the learner of Japanese, as well as by the fact that the learner was able to carry out the task of identifying reoccurring themes, despite limited previous experience in reading authentic Japanese texts.

A number of general usability issues were detected when Topics2Themes was used for analysing the Japanese texts. For instance, issues related to the functionality of presenting and resorting the information. These usability issues have been addressed, and a new version of the tool has been released. This new version of Topics2Themes will be used when the tool is applied to other text types.

5 Acknowledgements

The adaptation of Topics2Themes to Japanese, as well as the evaluation on Japanese texts, was funded by the Japan Society for the Promotion of Science, in the form of a "Postdoctoral Fellowships for Research in Japan (Short-term)" research grant.

The updated version of the tool, in which usability issues were corrected, was developed within the Språkbanken and SWE-CLARIN infrastructures, supported by the Swedish Research Council (2017-00626). Further developments and evaluations of the tool will also continue within the Språkbanken and SWE-CLARIN infrastructures.

We would also like to thank the reviewers, for their useful input, and the teachers at the IAY International Academy Japanese Language School in Sapporo, for making it possible to evaluate our tool.

References

Magnus Ahltorp. 2012. A personalizable reading aid for second language learners of Japanese. Master's thesis, Royal Institute of Technology, Sweden.

Eric Alexander, Joe Kohlmann, Robin Valenza, Michael Witmore, and Michael Gleicher. 2014. Serendip: Topic model-driven visual exploration of text corpora. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, VAST '14, pages 173–182. IEEE.

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.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, January.

David M. Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, April.

Guoray Cai, Feng Sun, and Yongzhong Sha. 2018. Interactive visualization for topic model curation. In *Proceedings of the ACM IUI 2018 Workshop on Exploratory Search and Interactive Data Analytics*, ESIDA '18. CEUR-WS.org.

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

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*, AVI '12, pages 74–77. ACM.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, KDD '96, pages 226–231. AAAI Press.

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. ACM.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent Dirichlet Allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211, June.

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. `https://ci.nii.ac.jp/naid/10027284215/en/`.

Daniel D. Lee and H. Sebastian Seung. 2001. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13*, NIPS '00, pages 556–562. MIT Press. Proceedings of the Neural Information Processing Systems Conference 2000.

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, June.

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*, 105:28–42, September.

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. ACL.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. `https://arxiv.org/abs/1301.3781`.

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

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, July.

Rafal Rzepka and Kenji Araki. 2017. What people say? Web-based casuistry for artificial morality experiments. In *Proceedings of the 10th International Conference on Artificial General Intelligence (AGI '17)*, volume 10414 of *LNCS*, pages 178–187. Springer International Publishing.

Maria Skeppstedt, Andreas Kerren, and Manfred Stede. 2018a. Vaccine hesitancy in discussion forums: Computer-assisted argument mining with topic models. *Studies in Health Technology and Informatics*, 247:366–370. Proceedings of the 29th Medical Informatics Europe Conference (MIE '18) — Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth.

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 3rd Workshop on Visualization as Added Value in the Development, Use and Evaluation of Language Resources at LREC '18*, VisLR III, pages 9–16. ELRA.

Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Closing the loop: User-centered design and evaluation of a human-in-the-loop topic modeling system. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, IUI '18, pages 293–304. ACM.

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*, ACL '05, pages 133–140. ACL.