Modeling Japanese Language Register Using Corpus Metadata

Bor Hodošček　(National Institute for Japanese Language and Linguistics) †

コーパスのメタデータを用いた日本語におけるレジスターのモデル化

ホドシチェク・ボル　（国立国語研究所）

1. Introduction

This study proposes to combine extra-linguistic and inter-linguistic features in the modeling of linguistic variation, or *register*, in Japanese. A comprehensive account of register variation in Japanese has recently become possible with the public availability of the Balanced Corpus of Contemporary Written Japanese (BCCWJ), a 100 million token corpus that contains a wide variety of written Japanese (Maekawa, 2007).

First, this study will provide necessary background on register studies and the connection between them and corpus metadata. Next, the metadata available in the BCCWJ are analyzed into several categories thought to influence register, taking into account any hierarchical properties that exist within the metadata. Finally, two pilot experiments using techniques from subgroup discovery (Wrobel, 1997; Langohr et al., 2013) and exceptional model mining (Duivesteijn, 2013) are conducted on the BCCWJ, showing some possibilities for further explorations of register variation within the BCCWJ.

2. Previous Research

2.1 Language Variation

Numerous studies on the topic of linguistic variation have been conducted using the different terminologies of style, genre, register, text type, and domain, to describe the variation observable in language (Eckert & Rickford, 2001; Biber, 1995; Irvine, 2001; Lee, 2001). A possible account of the differences between the terms is offered below (also see Lee, 2001):

- Register: A variety of language associated with the specific situation of use. Example: register of written academic Japanese; classroom conversation
- Genre: A category of language defined by a community, or associated with expected rhetorical structure and themes. Example: genre of Japanese research articles; crime novels
- Style: Variations in language associated with an individual's "unique" uses of language. Example: sensationalist style; vague written style
- Text type: A grouping of texts based purely on linguistic features. Example: informational text type
- Domain: Text devoted to a single topic or a small set of related topics, often inside one genre. Example: domain of computational linguistics

† bor.hodoscek@gmail.com

Different approaches to modeling language variation have been developed, mirroring the needs and interests of many fields, including: socio-linguistics, historical linguistics, dialect research, NLP, text classification, and authorship identification. However, for the purposes of the present work, we will mostly be concerned with register in the broader sense, covering the situation of use, community, individual unique usage of language–but all of which must be tied to actual linguistic features. The vocabulary used to describe register is furthermore limited to the metadata available within the BCCWJ, which is introduced in the following section.

2.2 Corpus Metadata

In the context of a corpus, we define metadata to be any data that describes some language-external fact about some part of a corpus. Therefore, this definition encompasses not only the common case of data describing some fact at the document level, but also includes data at the level of document groups, as well as data at the lower token, sentence, and paragraph levels (see Figure 1). A summary of the metadata annotations within the BCCWJ is provided in Maruyama (2012), while a further summary of metadata into formal and informal annotation approaches, as well as one based on granularity and dimension, is provided below through the lens of hierarchical classification systems available in the BCCWJ.

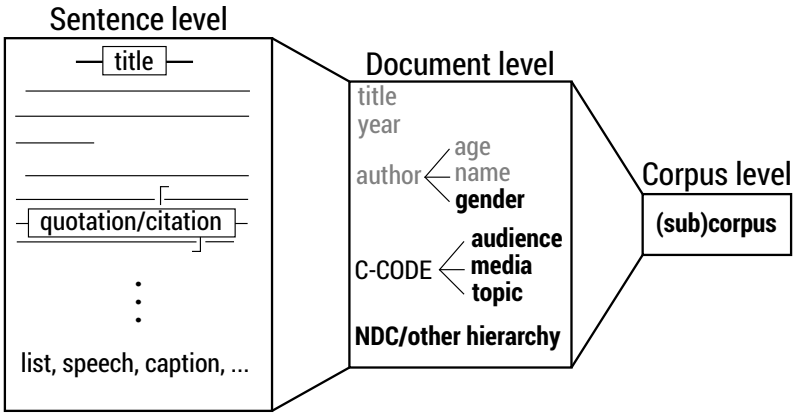

Figure 1: *Metadata at the sentence, document, and subcorpus levels. The metadata used in this study are marked in bold.*

Library classification systems such as the Dewey Decimal Classification (DDC) or the Japanese counterpart, the Nippon Decimal Classification (NDC), are classification systems which rely on expert catalogers and are arguably optimized for the efficient classification and storage of material in libraries. This guided approach to classification espouses a singular point of view and is thus susceptible to any oversights or gaps in knowledge of the catalogers. The organization of knowledge into ontologies is in many ways the most formal approach in this metadata category, and there are many cases where such a formal specification has been useful (Hirst, 2009, provides several).

In contrast, the other approach to metadata is more informal, and can involve collaborative classification of content using sets of tags in what is sometimes called a folksonomy (Vander Wal, 2007). This approach eschews the singular viewpoint (i.e., that of the compiler) of an ontology for the plurality of multiple viewpoints (i.e., that of many users). It is this property of the informal approach which

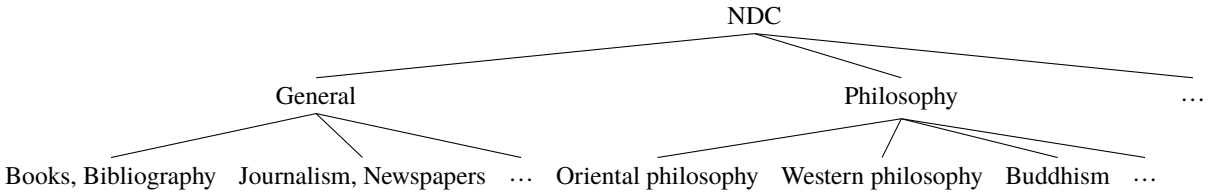

Figure 2: *A part of the hierarchical structure of the NDC.*

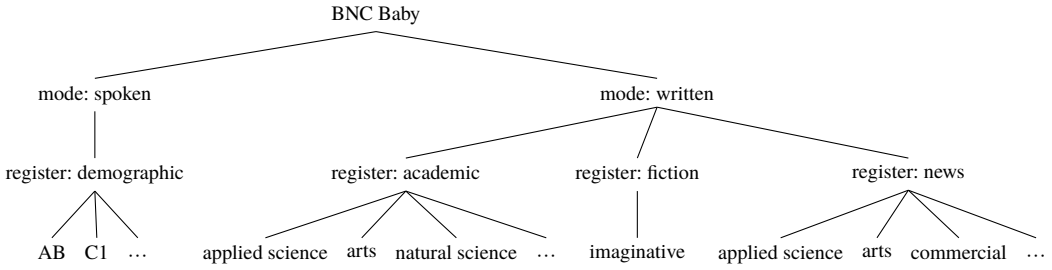

Figure 3: *A part of the hierarchical structure of the register-focused metadata within the British National Corpus Baby (as described in Gries, 2009, pp. 3–4).*

arguably makes it more flexible in describing a wider variety of metadata, given that the vocabulary of tags is unrestricted enough and that the number of taggers is sufficient to overcome incongruities in the tag data (Halpin, Robu, & Shepherd, 2007).

Finally, metadata can also be analyzed according to their granularity or dimensionality. The granularity of metadata refers to the complexity of structure contained in the metadata. For example, 1-dimensional metadata are composed of elements that are independent of one another[1]. An example of higher-dimensionality metadata is the NDC, in which 10 top-level categories further branch out into sub-categories (Figure 2).

As the annotation of corpus metadata depends on the research aims of the corpus and can thus encompass a diverse set of attributes, we limit the discussion to those corpus metadata which are involved in the identification of language variety of one kind or another. A representative example of corpus metadata that has been used for such purposes is that contained within the British National Corpus Baby (BNC Baby)[2]. For example, the text categories in the BNC Baby can be represented as a three-level hierarchy of mode, register, and sub-register, as shown in Figure 3.

Contrasting the hierarchy of the NDC with the register-focused hierarchy of the BNC Baby reveals limitations in using it as a proxy for register distinctions. A strict hierarchy such as the NDC is unable to capture all linguistic variation and situational factors. This is evident because the branching factor in the hierarchy, from parent to children, is often a combination of topic, register, and genre differences, which can interact at different levels. Instead, the realization of an ontology of linguistic variation with an associated metadata vocabulary would allow for the expression of arbitrary differences within and

[1]The Simple Dublin Core Metadata Element Set (DCMES) vocabulary is an example of this kind of linear structure (“Dublin Core Metadata Element Set, Version 1.1,” 2012).

[2]The complete list of BNC Baby categories is available from http://projects.oucs.ox.ac.uk/xaira/exercises/A1_xaira.html.

Table 1: *Number of tokens, sentences, paragraphs, and documents in the BCCWJ arranged according to media label. The last column "Hierarchy type" offers a short description of metadata at the media level.*

Code	Media label	Tokens	Sentences	Paragraphs	Documents	Hierarchy type
LB+PB+OB	Books	70,472,742	3,155,084	1,552,490	22,058	NDC, C-CODE
OY	Yahoo! Blogs	13,212,757	943,646	783,871	52,676	Yahoo! topics
OC	Yahoo! Q&A	12,088,127	780,510	624,616	91,445	Yahoo! topics
OM	Minutes of the Diet	5,600,649	139,802	45,810	159	place and type
OW	White papers	5,495,254	139,373	101,587	1,500	govt. division
PM	Magazines	5,114,752	281,765	155,260	1,996	magazine class.
OP	Local govt. newsletters	4,739,306	255,841	209,679	354	municipality
OL	Laws	1,206,481	33,289	25,364	346	subject
OT	School textbooks	1,126,214	63,370	45,952	412	subject and grade
PN	Newspapers	1,036,285	50,960	26,546	1,473	daily, evening
OV	Verse	237,685	18,974	18,974	252	haiku, etc.
	TOTAL	120,330,252	5,862,614	3,590,149	172,671	

between texts. However, the author has not yet encountered this kind of systematic encoding of the metadata inherent in corpora, and will, for the purposes of this study seek to examine only those parts of the metadata within the BCCWJ that more directly influence linguistic variation[3].

3. Materials

The BCCWJ contains a wide variety of contemporary written[4] Japanese documents with associated metadata of many forms. As can be seen in Table 1, the whole corpus, as used in this study, consists of approximately 120 million tokens[5], and is sub-divided according to general media labels. For the purposes of this study, the term "media label" is used in preference to the more established "subcorpus", as it provides a more accurate grouping of samples in the BCCWJ when one is not interested in differences between the three Books subcorpora (LB, PB, and OB).

Adding to the metadata that is unique to each media, the BCCWJ provides a variety of metadata for varying subsets of the corpus including: author name, gender, author decade of birth, publishing date, and publisher name. Most importantly for our purposes, up to 4 "genre" labels per document are provided, which allows the construction of hierarchies unique to each media label (see the "Hierarchy type" column in Table 1). It is quite evident that these "genre" labels represent different conceptualizations of genre, differing in their inclusion of topic, register, audience, and so on (see also Tanomura (2014) for another critique). The next section will detail the process of constructing hierarchies from

[3] See (Garbacz, 2006) for the beginnings of such an attempt based on the theory of document genres proposed in (Yates & Orlikowski, 1992; Orlikowski & Yates, 1994; Yates, Orlikowski, & Okamura, 1999).

[4] One notable exception is the Minutes of the Diet subcorpus, which consists of official transcriptions of dialogue.

[5] Tokens here refer to the Short Unit Words (SUWs) contained in the UniDic morphological dictionary version 2.1.2, which were extracted using the morphological analyzer MeCab version 0.966 from the C-XML variable length data of the BCCWJ, version 1.

these metadata.

4. Method

An understanding of the relationships between the metadata is a prerequisite to evaluating the discriminatory power of metadata in the modeling of language variation. Subgroup discovery, as described in the MIDOS algorithm (Wrobel, 1997), seeks to discover subgroups that have unusual distributional characteristics with respect to the entire population. Like MIDOS, we are looking for subgroups that have unusual distributional characteristics with respect to the entire population, but we are also further interested in unusual distributions with respect to the parent subgroup. First we define the language model that will be used to measure differences between different subgroups, and, in the succeeding subsection, detail the algorithm and modifications used to discover subgroups in the BCCWJ.

4.1 Language Model

The language contained in documents and subgroups is represented using a vector of word weights, calculated using a variant (wf.idf) of the well-known tf-idf (term frequency-inverse document frequency) formulation first proposed by Spärk Jones in 1972. Given a corpus containing documents, it is possible to weight words according to not only their frequency, but also to their dispersion across the number of documents they occur in. More formally:

$$\text{tf}(t, D) = \sum_{i=1}^{N} \text{number of occurrences of } t \text{ in } d_i$$

$$\text{df}(t, D) = \text{number of documents where } t \text{ occurs in}$$

$$\text{wf}(t, D) = 1 + \log_2 \text{tf}(t, D)$$

$$\text{idf}(t, D) = \log_2 \frac{N}{\text{df}(t, D)}$$

$$\text{wf.idf}(t, D) = \text{wf}(t, D) \cdot \text{idf}(t, D)$$

where t and d correspond to a term and a document, respectively. N is the total number of documents in the collection, which is 172,675 in the case of the BCCWJ. The term-frequency tf is weighted using sublinear scaling (wf), while the inverse document frequency idf is just the logarithm of the total number of documents divided by the number of documents containing t. This model, although simplistic, was chosen as a first step towards introducing more varied features that are able to capture more nuanced register variation in future work.

4.2 Subgroup Discovery and Exceptional Model Mining

Algorithm 1 gives the general outline of the SD/EMM process (a detailed explanation of the process is offered in Duivesteijn (2013), pp. 13–30). Intuitively, the algorithm starts the search from the vocabulary of features available (corresponding to the various metadata categories here). It then searches for increasingly small subsets of the data based on the most promising (according to the quality measure φ) subsets found so far. New subsets are thus comprised of conjoining the current promising subsets

(there are at most w of them) with new ones using the refinement operator η. For example, if a promising subset was (TOPIC = "politics"), the refinement operator would generate contrasting subgroups such as (TOPIC = "politics" $\wedge$ GENDER = "female") and (TOPIC = "politics" $\wedge$ GENDER != "female"). The quality score would then be computed by using the results of comparing the new subgroup and its complement with the whole dataset Ω. The search would continue until the maximum search depth (d) was reached, or until no more subgroups could be found, and the final result of which would be a list of the top-q subgroups sorted according to the chosen quality measure.

Algorithm 1: Beam search for top-q Exceptional Model Mining (reproduced from Duivesteijn, 2013, p. 19).

Input : Dataset Ω, QualityMeasure φ, RefinementOperator η, Integer w, d, q, Constraints C
Output: PriorityQueue resultSet

```
1  candidateQueue ← new Queue;
2  candidateQueue.enqueue({});
3  for Integer level ← 1; level ≤ d; level++ do
4      beam ← new PriorityQueue(w);
5      while candidateQueue ≠ ∅ do
6          seed ← candidateQueue.dequeue();
7          set ← η(seed);
8          foreach desc ∈ set do
9              quality ← φ(desc);
10             if desc.satisfiesAll(C) then
11                 resultSet.insert_with_priority(desc, quality);
12                 beam.insert_with_priority(desc, quality);
13             end
14         end
15     end
16     while beam ≠ ∅ do
17         candidateQueue.enqueue(beam.get_front_element());
18     end
19 end
20 return resultSet;
```

The following provides a detailed description of the parameters and methods used in Algorithm 1.

4.2.1 Distance function

The correlation between two wf.idf vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_m)$ is calculated using the sample correlation r:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

4.2.2 Beam-search parameters

The search width was set to 10, and the search depth was set to 8, though in practice no subgroups matching all the constraints were found beyond the fifth depth level.

4.2.3 Refinement operator

For nominal a_i with values $v_1, \ldots, v_g$ we add $\{D(a_i = v_j), D(a_i \neq v_j)\}_{j=1}^{g}$ to $\eta(D)$. The refinement of hierarchical data presents another challenge in that we must take into account the explicit constraints that hold within hierarchical data (cf. Park and Fürnkranz (2008), p. 2). Therefore, for each refinement of hierarchical metadata, all children of the current position in the hierarchy are generated.

4.2.4 Constraints

The minimum number of samples was chosen to be 20.

4.2.5 Quality measures

Following (Duivesteijn, 2013), two different quality measures were considered: φ_{scd} and φ_{ent}. The first quality measure, though statistically-oriented, revealed to be overly sensitive to the large subgroup document sizes the BCCWJ includes (which include over 170,000 documents) and measuring most subgroups' p value as zero. Thus the use of φ_{scd} was deemed to not be suitable for discovering subgroups in the present data, and other measures were considered:

- The most simple measure is simply the absolute difference between the correlation coefficient of the subgroup and its complement:

$$\varphi_{abs}(D) = |r^{G_D} - r^{G_D^C}|$$

- As the previous measure does not take into account the subgroup size or the distribution of the split of data between the subgroup and its complement, the entropy function φ_{ef} was considered:

$$\varphi_{ef}(D) = -{}^{n}/_{N}\, log_2\, {}^{n}/_{N} - {}^{n^C}/_{N}\, log_2\, {}^{n^C}/_{N}$$

- Finally, by combining φ_{abs} and φ_{ef}, we arrive at the heuristic measure φ_{ent}, which favors bigger, more balanced subgroups over those based on only the absolute difference between correlations:

$$\varphi_{ent}(D) = \varphi_{ef}(D) \cdot \varphi_{abs}(D)$$

5. Results

The results of two subgroup discovery runs conducted on the BCCWJ are presented in Table 2, both limited to the top 10 results for each run. The first run used φ_{ent} as the quality measure, while the second run used the unweighted φ_{abs} as the measure. All other parameters were kept constant during the two runs.

Table 2: *Results of top-q beam search using φ_{ent} and φ_{abs} as the quality measure (top 10 for each shown). Labels with no official English transcription translated by author.*

Quality measure ← φ_{ent} (Other parameters: q ← 40, beam width ← 10, search depth ← 8, minimum document coverage ← 20)							
Subgroup condition	φ_{ent}	φ_{abs}	φ_{ef}	r	r^C	G	G^C
CATEGORY="Yahoo! Q&A"	0.417	0.418	0.997	0.570	0.988	91,445	81,230
CATEGORY="Yahoo! Blogs"	0.293	0.330	0.887	0.647	0.977	52,680	119,995
TOPIC="literature and lit. criticism, etc." ∧ AUDIENCE="expert"	0.162	0.264	0.615	0.556	0.820	2,630	14,644
AUDIENCE="expert"	0.156	0.277	0.562	0.557	0.834	2,637	17,411
MEDIA="book" ∧ AUDIENCE="expert"	0.151	0.239	0.634	0.557	0.796	2,637	13,879
TOPIC="foreign literary novel" ∧ AUDIENCE="expert"	0.149	0.256	0.582	0.556	0.813	2,632	16,276
MEDIA="book" ∧ AUDIENCE="expert" ∧ TOPIC="medicine"	0.148	0.268	0.551	0.281	0.548	337	2,300
AUDIENCE="expert" ∧ TOPIC="medicine"	0.148	0.268	0.551	0.281	0.548	337	2,300
MEDIA="book" ∧ AUDIENCE="expert" ∧ TOPIC="elec. communication"	0.146	0.327	0.446	0.225	0.552	245	2,392
AUDIENCE="expert" ∧ TOPIC="elec. communication"	0.146	0.327	0.446	0.225	0.552	245	2,392
Quality measure ← φ_{abs} (Other parameters: q ← 40, beam width ← 10, search depth ← 8, minimum document coverage ← 20)							
Subgroup condition	φ_{abs}	φ_{ent}	φ_{ef}	r	r^C	G	G^C
CATEGORY="law"	0.833	0.017	0.021	0.166	0.999	346	172,329
TOPIC="foreign language"	0.753	0.012	0.016	0.113	0.866	29	20,019
MEDIA="comics" ∧ TOPIC="foreign language"	0.753	0.012	0.016	0.113	0.865	29	19,994
MEDIA="illustrated book" ∧ TOPIC="foreign language"	0.752	0.012	0.016	0.113	0.865	29	19,989
TOPIC="language learning"	0.735	0.013	0.017	0.130	0.866	32	20016
MEDIA="comics" ∧ TOPIC="language learning"	0.736	0.013	0.017	0.130	0.865	32	19,991
MEDIA="illustrated book" ∧ TOPIC="language learning"	0.735	0.013	0.017	0.130	0.865	32	19,986
CATEGORY="verse"	0.735	0.012	0.016	0.264	0.999	252	172,423
MEDIA="comics"	0.729	0.010	0.014	0.137	0.866	25	20,023
TOPIC="language learning" ∧ MEDIA="comics"	0.729	0.010	0.014	0.137	0.865	25	19,991

6. Discussion

Overall, the top scoring subgroup condition was CATEGORY, which is a joining of the media label with any hierarchical metadata that exist for it (see Table 1). While the Yahoo! Q&A and Blogs subcorpora displayed the largest differences in correlation to the global model when factoring in their relatively large size, the Law documents media displayed the greatest absolute difference. A perhaps more interesting subgroup condition found in these two lists is the AUDIENCE metadata extracted from the C-CODE, as it indeed describes an important element of register that is missing in otherwise more complete metadata such as the NDC hierarchy. The NDC or other media hierarchies did not feature strongly in the top results, though further examination of lower results did reveal more 3- and 4-part conjoined conditions.

7. Conclusion

Using both the metadata and the language data available within the BCCWJ, the present study conducted two pilot subgroup discovery experiments to uncover subgroups in the BCCWJ that exhibit divergent language usage based on the correlation differences between subgroups. However, the resulting subgroups from the pilot study revealed mostly those subsets of the data that were expected to exhibit specialized language, such as law documents or Internet data, with more complicated groupings of metadata not featuring prominently within the top results. Further experiments are needed to elucidate the relationships between the different metadata, both linear as well as hierarchical, within the BCCWJ.

8. Future Work

Future work should not only look to improve the subgroup discovery task or to improve feature extraction, but to also inform future annotation of corpus metadata, especially those metadata which help to uncover subgroups with divergent linguistic properties.

Finally, Wu, Markert, and Sharoff (2010) introduce a new method of quantifying genre hierarchies in terms of their visual and distributional imbalance based on tree balance entropy scores (pp. 755-757). Incorporating these measures could provide an additional quantification of subgroup distributionality alongside the entropy split function φ_{ef}, especially important in that it would permit comparisons between the different hierarchical metadata within the BCCWJ. These methods would also allow the hierarchies present in the BCCWJ to be compared to more established (from the point of view of suitability for register studies) metadata hierarchies such as those of the BNC Baby introduced in 2.2, or the Lancaster-Oslo/Bergen (LOB) Corpus[6].

Acknowledgements

This research was financially supported by JSPS Foreign Post-Doc Fellowship grant (no. P13303). A previous version of this paper was presented at the 19th Symposium of the Association for Database on the Humanities (Hodošček & Yamamoto, 2013).

References

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Dublin Core Metadata Element Set, Version 1.1. (2012, June 14). Retrieved October 30, 2013, from http://dublincore.org/documents/dces/

Duivesteijn, W. (2013). *Exceptional Model Mining* (Doctoral dissertation, Leiden Institute of Advanced Computer Science (LIACS), Faculty of Science, Leiden University).

Eckert, P. & Rickford, J. R. (2001). *Style and sociolinguistic variation*. Cambridge University Press.

Garbacz, P. (2006). An outline of a formal ontology of genres. In *Knowledge science, engineering and management* (pp. 151–163). Springer.

Gries, S. T. (2009). Bigrams in registers, domains, and varieties: a bigram gravity approach to the homogeneity of corpora. In *Proceedings of Corpus Linguistics 2009*. University of Liverpool.

Halpin, H., Robu, V., & Shepherd, H. (2007). The complex dynamics of collaborative tagging. In *Proceedings of the 16th International Conference on the World Wide Web* (pp. 211–220). WWW '07. Banff, Alberta, Canada: ACM. doi:10.1145/1242572.1242602

Hirst, G. (2009). Ontology and the lexicon. In *Handbook on ontologies* (pp. 269–292). Springer.

Hodošček, B. & Yamamoto, H. (2013). *The role of metadata in the Balanced Corpus of Contemporary Written Japanese in the analysis of register*. Presented at the 19th Symposium of the Association for Database on the Humanities, Ritsumeikan University.

[6] A list of categories used in the LOB corpus is available from http://icame.uib.no/lob/lob-dir.htm.

Irvine, J. T. (2001). “Style” as distinctiveness: the culture and ideology of linguistic differentiation. In P. Eckert & J. R. Rickford (Eds.), *Style and sociolinguistic variation* (pp. 21–43). Cambridge University Press.

Langohr, L., Podpečan, V., Petek, M., Mozetič, I., Gruden, K., Lavrač, N., & Toivonen, H. (2013). Contrasting subgroup discovery. *The Computer Journal*, *56*(3), 289–303.

Lee, D. Y. W. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, *5*(3), 37–72. Retrieved from http://llt.msu.edu/vol5num3/pdf/lee.pdf

Maekawa, K. (2007). KOTONOHA and BCCWJ: Development of a Balanced Corpus of Contemporary Written Japanese. In *Proceedings of the first international conference on Korean language, literature, and culture* (Vol. 2, pp. 158–177). Corpora and Language Research. Seoul.

Maruyama, T. (2012, March). The role of metadata in the analysis of large-scale corpora. In *Dai sankai kōpasu nihongo wākushoppu yokōshū [Proceedings of the 3rd Workshop on Japanese Corpus Linguistics]* (pp. 203–210). Dai sankai nihongo kōpasu wākushoppu [3rd Workshop on Japanese Corpus Linguistics]. Tokyo, Japan.

Orlikowski, W. J. & Yates, J. (1994). Genre repertoire: the structuring of communicative practices in organizations. *Administrative science quarterly*, 541–574.

Park, S.-H. & Fürnkranz, J. (2008). Multi-label classification with label constraints. In *Proceedings of the ECML PKDD 2008 Workshop on Preference Learning (PL 2008)* (pp. 157–171). Antwerp, Belgium.

Spärk Jones, K. (1972). A statistical interpretation of term importance in automatic indexing. *Journal of Documentation*, *28*(1), 11–21.

Tanomura, T. (2014). BCCWJ no shiryōteki tokusei: kōpasu rikai no jūyōsei [On the characteristics of the BCCWJ as material: The importance of understanding the corpus]. In *Kōpasu to nihongogaku [Corpus and Japanese linguistics]* (Vol. 6, Vols. 8). Kōza nihongo kōpasu [Corpus textbook series]. Asakura Publishing Co., Ltd.

Vander Wal, T. (2007). Folksonomy. Retrieved October 30, 2013, from http://vanderwal.net/folksonomy.html

Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. In *Principles of Data Mining and Knowledge Discovery* (pp. 78–87). Springer.

Wu, Z., Markert, K., & Sharoff, S. (2010). Fine-grained genre classification using structural learning algorithms. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 749–759). Association for Computational Linguistics.

Yates, J. & Orlikowski, W. J. (1992). Genres of organizational communication: a structurational approach to studying communication and media. *Academy of management review*, *17*(2), 299–326.

Yates, J., Orlikowski, W. J., & Okamura, K. (1999). Explicit and implicit structuring of genres in electronic communication: reinforcement and change of social interaction. *Organization science*, *10*(1), 83–103.