Analyzing the characteristics of academic paper categories by using an index of representativeness ⋆

Takafumi Suzuki[a], Kiyoko Uchiyama[b], Ryota Tomisaka[c], and Akiko Aizawa[b]

[a] Faculity of Sociology, Toyo University,
5-28-20, Hakusan, Bunkyo-ku, Tokyo 112-8606, Japan
takafumi_s@toyo.jp
[b] Digital Content and Media Sciences Research Division, National Institute of Informatics
2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
{kiyoko, aizawa}@nii.ac.jp
[c] Graduate School of Information Science and Technology, The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo 112-0023, Japan
ryota.tomisaka@gmail.com

Abstract. This study proposes an index of representativeness for analyzing the characteristics of academic paper categories. Many textual indices have been proposed in the field of computational stylistics, but all of the previous indices are limited in that they (a) focus only on the styles of the texts; (b) return an absolute value for every text, and (c) are based on the number of tokens. In this study, we propose an index of representativeness that does not have the weaknesses of the previous indices. Our index is based on the h-index that was originally proposed in the field of scientometrics. We redefine it here for textual data. We show the effectiveness of our index for analyzing the characteristics that differ between four genres and three subfields in Japanese academic papers.

Keywords: academic papers, h-index, lexical indicators

1 Introduction

Many textual indices have been proposed in the field of computational stylistics. The number of tokens, i.e., the frequency of words in a document, is the simplest. Some indices use information related to words that appear only one or two times in the text, while others take the entire frequency spectrum into account (Kageura, 2000; Tweedie and Baayen, 1998). Textual indices have been studied for various applications such as authorship analysis, genre detection, and chronological text mining (Argamon *et al.*, 2007; Koppel *et al.*, 2009; Stamatatos, 2009), but they have three flaws in common: (a) they shed light only on the styles of text; (b) they return an absolute value, meaning that they do not consider the relative values for different keyword sets or the relative position of a document in the document set; and (c) they are based on the number of tokens.

Our problem is the use of textual indices to analyze the characteristics of the categories of academic papers. Since a wide variety of texts are available, these types of the 'characterization' problems, as well as the 'classification' problems, are currently increasing in importance. To deal with the characterization problems, we aim at deriving a meaningful level of knowledge

⋆ This study was supported by Grant-in-Aid for Scientific Research 21800087 for Young Scientists (Start-up) and 23700288 for Young Scientists (B), from the Ministry of Education, Culture, Sports, Sciences and Technology, Japan, and Grant for joint research from National Institute of Informatics. This study used the text data provided by the Information Processing Society of Japan (IPSJ). Earlier version of this study was presented at 16th Annual Meeting of the Association for Natural Language Processing (NLP2010). We would like to thank all of them for helping us conducting and improving this research.

from the textual characteristics using the summarized indices given to the categories (Kageura and Abekawa, 2009; Suzuki, 2009). Our categories are stylistic (genre) as well as content-dependent (subfield); thus, our index must be capable of analyzing both of them. The categories we deal with are similar to each other but different from the conventional ones used in genre detection or topic classification. Thus, it is desirable for our index to be able to deal with the relative position of a document in a document set. Token number-based indices have a common problem in that they are affected by several frequent words. This means we should create indices that do not directly use a token number that will also require less computational effort.

Against this background, we propose an index of representativeness that overcomes the drawbacks of the previously used indices. Our index is based on the h-index that was originally proposed in the field of scientometrics (Hirsch, 2005). After transforming this index for textual data, we use it for measuring how well a document represents the entire document set regarding several perspectives determined by keyword lists. It can analyze stylistic categories as well as content-dependent ones; our index uses content words (keyword lists); that means it can be used for analyzing content-dependent categories. Furthermore, it is based on the rank-type information, instead of the token-number information; that means it can be used for analyzing genre-based categories, because the rank-type information is usually considered to represent styles (Tweedie and Baayen, 1998). Our index captures both stylistic and content-dependent characteristics (Suzuki and Kageura, 2007) based on this information. In addition, it can use different keyword lists and takes into consideration the relative position of a document in a document set; and it also uses the ranks of the types, i.e., the ranked frequencies of different words (keywords) appearing in a document. We assessed the effectiveness of our index in analyzing the different characteristics of four genres and three subfields in Japanese academic papers.

The rest of this paper is organized as follows. We first explain the original h-index and its characteristics in Section 2. In Section 3, we explain the proposed method, including the redefinition of the index for textual data. We explain the experimental setup in Section 4 and discuss the results in Section 5. We review the related work in Section 6 and conclude with final remarks in Section 7.

2 *h*-index

The h-index was first proposed in the field of scientometrics by (Hirsch, 2005) and is defined as follows.

> A scientist has index h if h of his or her X papers each have at least h citations and the other $(X - h)$ papers each have $\leq h$ citations.

For example, a scientist whose h is 30 has published 30 papers, each of which has at least 30 citations.

This index has many advantages over other indices. Other possible indices and their limitations are as follows: (a) the total number of papers is not a measure of the importance or impact of papers; (b) the total number of citations can be inflated by a small number of highly cited papers; (c) the citations per paper rewards low productivity papers while it penalizes high productivity papers; (d) the number of significant papers is arbitrary as to the 'significance' criteria; and (e) the number of citations for each of the q most-cited papers is not a single number and is more complicated to obtain and compare. The h-index avoids all the drawbacks of these indices, yet it can measure the broad impact of scientists' productivity. That is why it is one of the most frequently used indices for evaluating the scientific research. It was pointed out that the h-index can be defined and used for any data that follows a power law (Egghe and Rousseau, 2006b). Here, we shall redefine it for textual data. There are many modifications and sophistications associated with this index for better interpreting the phenomena in scientometrics (Egghe and Rousseau, 2006a; Sidiropoulos *et al.*, 2007), but most of these modifications and sophistications

do not directly lead to better interpreting the textual data from our point of view, thus we added our own modifications to this index.

3 Proposed method

First, we briefly describe our method and our redefinition of the h-index. After that, we explain the keyword lists used in the experiment.

3.1 Overview

We analyze the characteristics of categories according to the following process.

(i.) Given a target category (a set of documents),

(ii.) extract a set of words using the keyword lists,[1]

(iii.) collect the document frequencies ($DF(w, d)$) of (ii) in (i),

(iv.) using (iii), calculate the t/b ratio (see below) for each document in a category.

(v.) The category index is the median and the iqr value of (iv).

3.1.1 Redefinition of h-index for textual data Our basic idea is to deal with the scientometric data and textual data from the integrated viewpoint of quantitative informatics. Although making a reference to a paper and using a word in a paper seems to be a different activity, many previous studies have indicated that the mathematical and quantitative characteristics shown in these two phenomena were very similar to each other.[2] Accordingly, we can define the relations between these two types of data as follows.

- researchers: documents,
- number of publications: number of word types included in the document,
- number of citations: number of documents including the word types, i.e., document frequencies ($DF(w, d)$).

The redefined h-index for the textual data is basically as follows.

> A document has an index h if h of its X word types each have $DF(w, d) \geq h$ and the other $(X - h)$ word types each have $DF(w, d) \leq h$.

This index, based on the information about the *ranks* of the *types* instead of the *number* of *tokens*, reflects how general words (words with high $DF(w, d)$) are used in the texts. It is an index of the 'representativeness' of the text in a document set. A high h-value means that general words are frequently used in the texts, whereas a low h-value means that general words are infrequently used in the texts. This index is superior to indices (a) - (e) mentioned in Section 2 because it succeeds the characteristics of the original h-index.

3.1.2 Scaling The number of types and $DF(w, d)$ are quite different in scale from the number of publications and the number of citations; thus, we shall use the following scaling.[3]

$$DF'(w, d) = \lfloor DF(w, d) \times \alpha \times \frac{V_N(d)}{max_{w \in \omega}(DF(w, d))} \rfloor$$

[1] In this paper, we used keyword lists for extracting a set of words, but part-of-speech tags or automatic term extraction can also be used according to the purposes and target data of the researches.

[2] We will investigate the reason for this in future. See also (Egghe, 2009; Kageura, 2000).

[3] The number of publications of an author will be at most 10^1 (yearly) to scale and the number of citations of a paper will be at most 10^{10} (yearly) to scale, but the number of types and DF will have larger differences in many cases.

Table 1: Basic data of our corpora

	IPSJ			author		
		$V_N(d)$			$V_N(d)$	
	doc	median	iqr	doc	median	iqr
JNL	2452	91	31.00	2753	220	80.00
MAG	1194	75	41.00	1212	201	100.25
SIG	6751	63	29.00	7333	152	76.00
TRANS	933	96	31.00	1035	214	75.50
CS	2113	65	33.00	2328	141	120.25
IE	2856	61	24.00	3108	155	66.00
FR	1782	66	30.00	1897	162	77.00

where $max_{w\in\omega}(DF(w,d))$ represents the maximum $DF(w,d)$ value of the keyword w in keyword lists ω for every document, and $V_N(d)$ represents the number of types, i.e., the number of different keywords appearing in the document.

The scaling factors adjust the difference in scale between $V_N(d)$ and $DF(w,d)$. In particular, the α parameter determines the best ratio between $V_N(d)$ and $DF(w,d)$ for every used dataset; a higher (lower) α means that a larger (smaller) number of types are included in h. We select the α that gives the highest inter-quartile range (iqr), from 0.1 to 3.0 incremented by 0.1, because large variances between documents are desirable for our purposes. Moreover, by observing the stability of the results after changing α, we can confirm the reliability of the results.[4] The final definition of our h-index for textual data is as follows. We call it h' in what follows.

> A document has index h' if h' of its X word types each have $DF'(w,d) \geq h'$ and the other $(X - h')$ word types each have $DF'(w,d) \leq h'$ each.

3.1.3 Three indices based on the h'-index We shall examine the following three kinds of h'-indices.

Top h'-index This index is exactly the same as mentioned above. We name it the top h'-index simply to distinguish it from the bottom h'-index. This index reflects how general words (ones with high $DF(w,d)$) are used in text.

Bottom h'-index This index substitutes $DF''(w,d)$ for $DF'(w,d)$ when calculating h'.

$$DF''(w,d) = \lfloor DF^*(w,d) \times \alpha \times \frac{V_N(d)}{max_{w\in\omega}(DF^*(w,d))} \rfloor$$

where

$$DF^*(w,d) = max_{w\in\omega}(DF(w,d)) + 1 - DF(w,d)$$

This is a simple substitution to change high $DF(w,d)$ values to low $DF(w,d)$ values, and vice versa. This index, as opposed to the top h' index, is a measure of how special words (ones with low $DF(w,d)$) are used in texts.

[4] It can be used as a sort of significant test.

t/b ratio The third index is the ratio between the top h' and the bottom h':

$$t/b\ ratio = \frac{top\ h'}{bottom\ h' + 1}$$

This index calculates the ratio of the top h'-index to the bottom h'-index. A high t/b ratio means that general words (high $DF(w, d)$) are more frequently used (in the level of types) than special words (low $DF(w, d)$), and a low t/b ratio means that special words are more frequently used. The top h' and bottom h' heavily depend on the number of word types. Due to this, we shall use the t/b ratio for analyzing the characteristics of the genres and subfields. The value of this index is also independent of the number of documents used in the experiments. Their variances of this index should be smaller along with the increasing number of documents, but it has no effects on our interpretation of the results.

3.2 Keyword lists

One of the privileges of our index is the calculability for different keyword lists. It is not necessary to use several keyword lists for our methods, but by selecting two different (or opposite) types of keywords, we will strengthen our discussion of the results. In this study, we used keyword lists that included (a) general and (b) specific keywords in the field of information sciences, and then we can effectively analyze the characteristics of four genres and three subfields in Japanese academic papers.

More precisely, we extracted keywords from the data by using two keyword lists: (a) the IPSJ handbook index list (Joho Syori Gakkai, 1995) and (b) the author index list. List (a) is an index that includes 5,958 basic, low-expertise terms used in information sciences. List (b) is an index list that includes 16,935 terms that authors themselves added to their papers. While (a) includes basic conceptual keywords used in information sciences, it also (b) includes high expertise keywords based on the authors' selection. List (b) also includes numbers, mathematical expressions, and symbols. Both of them include high frequency words as well as low frequency words, and include terminology words as well as rather general words. After we applied morphological analysis to the data and keyword lists, we extracted the keywords by matching. Note that we used a Japanese morphological analysis system called MeCab[5] for this purpose.

We use these two types of keywords for calculating the h'-index in order to shed light on the different characteristics of the genres and subfields; i.e., the t/b ratio using IPSJ keywords should show the representativeness for basic, low-expertise keywords, and the t/b ratio using the author keywords should show the representativeness of special, high-expertise keywords. In so doing, we can obtain meaningful knowledge for understanding the characteristics of the four genres and three subfields.[6]

4 Experimental setup

The experiments used the text data provided by the Information Processing Society of Japan (IPSJ). This data consists of four kinds of published material, i.e., journals (JNL), magazines (MAG), technical reports for special interest groups (SIG), and transactions (TRANS). The JNL consists of standard journal papers, while the SIG consists of the reports from 39 special interest groups. The TRANS is a new journal with fewer topics compared to the JNL.[7] The most signif-

[5] mecab.sourceforge.net

[6] When we incidentally calculated our index by using the keyword lists, that index separates English texts from Japanese ones quite well.

[7] In fact, JNL and TRANS are not clearly distinguished from each other in terms of the communicative purpose, audience and form. We distinguish between them in this paper because it is in the original distributed data, and we thought it is better to distinguish them for further use in our findings and discussion. It should be noted that our main findings and discussion remain true even if we merge the two categories.

Table 2: h'-index using IPSJ/author keywords

	top		bottom		t/b ratio	
IPSJ	median	iqr	median	iqr	median	iqr
JNL	31	11.00	52	25.00	0.574	0.117
MAG	30	10.00	49	24.00	0.591	0.122
SIG	29	11.00	49	25.00	0.567	0.122
TRANS	40	11.00	73	30.00	0.531	0.097
	top		bottom		t/b ratio	
author	median	iqr	median	iqr	median	iqr
JNL	95	17.00	202	57.00	0.469	0.065
MAG	85	20.00	201	70.00	0.412	0.051
SIG	80	11.00	167	37.00	0.472	0.067
TRANS	95	17.00	196	55.25	0.481	0.063

icant difference is that of the MAG from the others. The MAG consists of book reports, event reports, and introductory essays, while the other genres basically consist of research papers that include original research findings. Thus, the MAG should be well represented by ordinary, i.e., low-expertise, keywords, and not by professional, i.e., high-expertise, keywords.

The 39 groups of SIGs are classified into three large groups, i.e., Computer Sciences (CS), Information Environment (IE), and Frontier (FR). The CS deals with the basic and theoretical issues in the information sciences. Its papers are related to algorithms, programming, and operating systems. They include many symbols and mathematical expressions and use a lot of professional keywords. The IE deals with conventional applications such as information systems, human-computer interaction, and computer graphics. The FR includes new and various applications such as game informatics, music and computers, and computers and humanities. These papers deal with many new concepts and have many new keywords that are not likely to be of the professional sort.

We used texts including more than the median number of word types, i.e., 72 types for IPSJ keywords experiments and 172 types for author keywords experiments, in order to exclude noisy data with corrupted characters and characters in other languages.[8] Table 1 lists the number of documents[9] and the median and iqr of $V_N(d)$ for each category.

5 Results and discussion

Section 5.1 discusses the results for the four genres and three subfields, and Section 5.2 discusses the results from the empirical examination of the characteristics of the indices.

5.1 Experimental results

5.1.1 Results for JNL, MAG, SIG, and TRANS genres We calculated the h'-indices for each text. Table 2 lists the median and iqr values of the top h', bottom h', and t/b ratio for the JNL, MAG, SIG, and TRANS categories. As will be shown in Section 5.2.2., $\alpha = 1.9$ had the highest iqr (0.120) regarding the t/b ratio for the IPSJ keywords, whereas $\alpha = 1.8$ (iqr, 0.072) had the highest iqr (0.120) for the author keywords. Thus we selected these α values to create Table 2.

[8] After trying other various types of criteria, we decided to do this because it was the best way to exclude noisy data from our corpora. For applying our index to other data, this process is not needed. In the future, we will develop the methods to automatically exclude these types of noisy data.

[9] More precisely, the 'doc' value represents the number of documents including at least one (IPSJ/author) keyword. According to this, the 'doc' values are different between experiments using IPSJ and those using author keywords in Table1, but the experiments themselves were basically conducted using the same document set.

The upper right column of Table 2 shows that MAG has a much higher t/b ratio than the other three genres for the IPSJ keywords, while the lower right column shows that it has a much lower t/b ratio for the author keywords. A Wilcoxson rank sum test showed a significant difference between MAG and the other genres regarding the t/b ratio ($p < .01$).

These results clearly reflect the special characteristics of the MAG genre. MAG consists of book reports, event reports, and introductory essays; thus, its documents should indeed be represented by low-expertise keywords, rather than by high-eaxpertise keywords. These results clearly shed light on the microscopic differences between MAG and the other genres.

Table 3: h'-index using IPSJ/author keywords

	top		bottom		t/b ratio	
IPSJ	median	iqr	median	iqr	median	iqr
CS	27	12.50	46	28.00	0.569	0.133
IE	30	11.00	51	25.00	0.569	0.118
FR	30	11.00	51	25.00	0.561	0.112
	top		bottom		t/b ratio	
author	median	iqr	median	iqr	median	iqr
CS	82	9.00	162	30.25	0.497	0.058
IE	80	12.00	166	35.00	0.470	0.066
FR	80	13.00	172	40.00	0.457	0.060

5.1.2 Results for CS, IE, and FR subfields We calculated the h'-index for every text. Table 3 lists the median and iqr values of the top h', bottom h', and t/b ratio for the CS, IE, and FR subfields. The same parameter (α) values as mentioned above were selected.

The right column in Table 3 shows that the t/b ratio of FR is lower than that of the other two categories (CS/IE) for both the IPSJ keywords (upper columns) and author keywords (lower columns). It also shows that the t/b ratio of CS is higher than that of the other two categories for the author keywords. A Wilcoxson rank sum test revealed the significant differences regarding the t/b ratio between FR and the others and CS and the others for the author keywords ($p < .01$).

These results reflect the characteristics of the CS and FR subfields. The CS includes theoretical studies; thus, it is reasonable to say that it is represented by high-expertise keywords. FR deals with the frontier of information sciences, and it includes many new topics. Thus, it is reasonable to say that it is 'not' represented by high-expertise keywords.

The results using IPSJ keywords are not stable as will be shown in Section 5.2.2., however, the results are consistent with the characteristics of the three subfields, CS and IE include standard and traditional subfields, thus they are represented by low-expertise, but standard keywords in information sciences, but FR includes many new topics, thus they are not represented by these types of standard keywords either. In other words, FR includes various topics, thus it is not represented by any type of keywords lists.

5.2 Characteristics of the indices

5.2.1 Dependency of the t/b ratio on the number of types In order to confirm the dependency of these indices on the number of types, we calculated the Spearman rank coefficient. The ρs between the top, bottom h'-index, the t/b ratio, and the number of types are respectively 0.086, 0.076, and -0.019 for the IPSJ keywords and 0.881, 0.997, and -0.643 for the author keywords. These results show that the t/b ratio decreased the number of type dependencies relative to those of the top and bottom h'-indices. These results indicate that there is almost no correlation between the t/b ratio and the number of types especially for the IPSJ keywords.

5.2.2 Stability of the results for changing α Figure 1 shows the relation between α and the median of the t/b ratio. The figures show that the ranks of the four genres and three subfields are stable, except in the experiment on the three subfields using the IPSJ keywords. The stability of these results can be used for testing the reliability of the results.

6 Related work

There are many studies on textual indices. They were systematically surveyed by (Tweedie and Baayen, 1998). Textual indices have been used for judging the readability of texts (Tuldava, 1993) and detecting the authors or genres of text (Kenny, 1982; Jin and Murakami, 2003).

Lexical frequency profiles were first proposed in (Laufer and Nation, 1995), and in studies in (Laufer and Nation, 1999; Hu and Nation, 2000; Laufer, 2005; Meara, 2004; Meara, 2005; Meara, 2006; Muncie, 2002; Edwards and Collins, 2010). Although it was produced for different purposes and used mainly in English, their discussion will be useful for the scholars that are interested in the character of textual indices.

While most textual indices focus only on the styles of the text and return absolute, token based values, our index using the ranks of the frequencies of the 'types' of text enable us to analyze the microscopic differences between content-dependent categories as well as the differences between stylistic categories.

The h-index was first proposed by (Hirsch, 2005), and there have been many studies that have investigated the mathematical characteristics and empirical usefulness of this and related indices (Antonakis and Lalive, 2008; Costas and Bordons, 2007; Egghe *et al.*, 2009; Egghe and Rousseau, 2006a). Our study is the first to redefine this index for textual data.

Hisamitsu *et al.* (2000) deals with the concept of 'representativeness', but it defines the representativeness for keyword extraction. Our study defines this concept for documents.

7 Conclusion

This study proposed an index of the representativeness for analyzing the characteristics of the genres and subfields of Japanese academic papers. We conducted an experiment that showed our 'representativeness' index was very useful for analyzing the microscopic differences between the characteristics of the four genres and three subfields.

Our index was in fact slightly dependent on the number of types. Within the scope of our paper, however, major changes for our interpretation were not required. We shall investigate the effects systematically in the future.

Our resources and findings will be useful for academic mining (Ichise *et al.*, 2007), which is a growing area of artificial intelligence research, and our method for analyzing the microscopic differences in the categories such as the genres and subfields will be useful for adapting techniques from one text domain to another text domain (domain adaptation). In the future, we will systematically enumerate such categories and provide a clearer perspective on the problem.

References

Antonakis, John and Rafael Lalive. 2008. Quantifying scholarly impact: IQp versus the Hirsch h. *Journal of the American Society for Information Science and Technology*, 59(6), 956–969.

Argamon, Shlomo, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6), 802–822.

Costas, Rodrigo and María Bordons. 2007. The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics*, 1(3), 193–203.

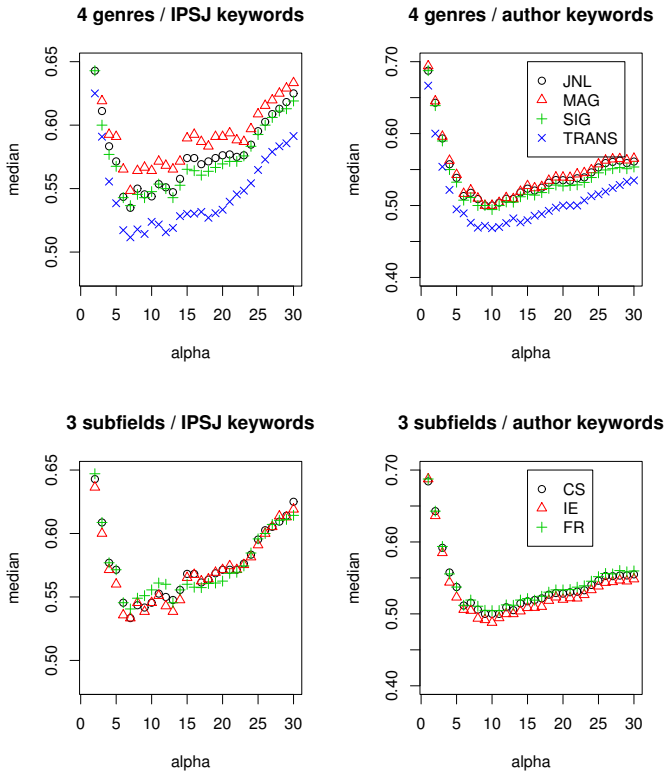

Figure 1: Relation between α and median of t/b ratio

Edwards, Roderick and Laura Collins. 2010. Lexical frequency profiles and Zipf's law. *Language Learning*.

Egghe, Leo. 2009. Performance and its relation with productivity in lotkaian systems. *Scientometrics*, 81(2), 567–585.

Egghe, Leo, Liming Liang, and Ronald Rousseau. 2009. A relation between h-index and impact factor in the power-law model. *Journal of the American Society for Information Science and Technology*, 60(11), 2362–2365.

Egghe, Leo and R. Rousseau. 2006a. Theory and practise of the g-index. *Scientometrics*, 69(1), 131–152.

Egghe, Leo and Ronald Rousseau. 2006b. An informetric model for the Hirsch-index. *Scientometrics*, 69(1), 121–129.

Hirsch, Jorge E. 2005. An index to quantify an individual's scientific research output. In *Proceedings of the National Academy of Science of the United States of America*, volume 102, pp. 16569–16572.

Hisamitsu, Toru, Yoshiki Niwa, Shingo Nishioka, Hirofumi Sakurai, Osamu Imaichi, Makoto Iwayama, and Akihiko Takano. 2000. Extracting terms by a combination of term freqency and a measure of term representativeness. *Terminology*, 6(2), 211–232.

Hu, Marcella Hsueh-chao and Paul Nation. 2000. Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403–430.

Ichise, Ryutaro, Setsu Fujita, Taichi Muraki, and Hideaki Takeda. 2007. Research mining using the relationships among authors, topics and papers. In *IV07: 11th International Conference Information Visualization*, pp. 425–430.

Jin, Mingzhe and Masakatsu Murakami. 2003. Bunsyo no tokei bunseki to wa. In Syunichi Amari, K Takeuchi, A Takemura, and Yukito Iba, eds., *Gengo to Shinri no Tokei: Kotoba to Kodo no Kakuritsu Moderu ni yoru Bunseki*, pp. 3–57. Iwanami Syoten, Tokyo.

Joho Syori Gakkai, ed. 1995. *Sinban Joho Syori Handobuku*. Ohmsha, Tokyo.

Kageura, Kyo. 2000. *Keiryou Johogaku*. Maruzen, Tokyo.

Kageura, Kyo and Takeshi Abekawa. 2009. Nlp meets library science: providing a set of enhanced language reference tools for online translators. In *Proceedings of A-LIEP2009: Asia-Pacific Conference on Library & Information Education and Practice*, pp. 538–547.

Kenny, Anthony. 1982. *The Computation of Style: An Introduction to Statistics for Students of Literature and Humanities*. Pergamon Press, Oxford.

Koppel, Moshe, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60.

Laufer, Batia. 2005. Lexical frequency profiles: from monte carlo to the real world, a response to Meara (2005). *Applied Linguistics*, 26(4), 582–588.

Laufer, Batia and Paul Nation. 1999. A vocabulary-size test of controlled productive ability. *Language Testing*, 16(1), 33–51.

Laufer, Betia and Paul Nation. 1995. Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics*, 16(3), 307–322.

Meara, Paul. 2004. Modelling vocabulary loss. *Applied Linguistics*, 25(2), 137–155.

Meara, Paul. 2005. Lexical frequency profiles: a monte carlo analysis. *Applied Linguistics*, 26(1), 32–47.

Meara, Paul. 2006. Emergent properties of multilingual lexicons. *Applied Linguistics*, 27(4), 620–644.

Muncie, James. 2002. Process writing and vocabulary development: comparing Lexical Frequency Profiles across drafts. *System*, 30, 225–235.

Sidiropoulos, Antonis, Dimitrios Katsaros, and Yannis. 2007. Generalized hirsch *h*-index for disclosing latent facts in citation networks. *Scientometrics*, 72, 253–280.

Stamatatos, Efstathios. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3), 538–556.

Suzuki, Takafumi. 2009. Extracting speaker-specific functional expressions from political speeches using random forests in order to investigate speakers' political styles. *Journal of the American Society for Information Science and Technology*, 60(8), 1596–1606.

Suzuki, Takafumi and Kyo Kageura. 2007. Exploring the microscopic characteristics of Japanese prime ministers' Diet addresses by measuring the quantity and diversity of nouns. In *PACLIC21: Proceedings of the 21th Pacific Asia Conference on Language, Information and Computation*, pp. 459–470.

Tuldava, Juhan. 1993. The statistical structure of a text and its readability. In *Quantitative Text Analysis*, pp. 215–227. Wissenschaftlicher Verlag Trier, Trier.

Tweedie, Fiona J. and R. Harald Baayen. 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32, 323–352.