Temporal Aspects of Terminology for Automatic Term Recognition: Case Study on Women's Studies Terms

Junko Kubo, Keita Tsuji, Shigeo Sugimoto
Graduate School of Library, Information and Media Studies, University of Tsukuba
1-2, Kasuga, Tsukuba, Ibaraki, 305-8550, Japan
E-mail: kuboj35@slis.tsukuba.ac.jp, keita@slis.tsukuba.ac.jp, sugimoto@slis.tsukuba.ac.jp

Abstract

The purpose of this paper is to clarify the temporal aspect of terminology focusing on the dictionary's impact on terms. We used women's studies terms as data and examined the changes of their values of five automatic term recognition (ATR) measures before and after dictionary publication. The changes of precision and recall of extraction based on these measures were also examined. The measures are TFIDF, C-value, MC-value, Nakagawa's FLR, and simple document frequencies. We found that being listed in dictionaries gives longevity to terms and prevent them from losing termhood that are represented by these ATR measures. The peripheral or relatively less important terms are more likely to be influenced by dictionaries and their termhood increase after being listed in dictionaries. Among the termhood, the potential of word formation that can be measured by Nakagawa's FLR seemed to be influenced most and the terms gradually gained it after being listed in dictionaries.

1. Introduction

Terminologies change over time because of many factors. The influential factors on such changes are: (1) publication of papers which contain innovative findings, (2) publication of standard dictionaries and textbooks, (3) establishment of academic societies, (4) reports in the media such as newspapers, and (5) other linguistic (phonetic, semantic, etc.) factors. While many methods have been proposed for automatic term recognition (henceforth ATR), little attention has been paid to these factors. If these temporal and social aspects of terminology are taken into consideration and are incorporated in ATR, its performance can be improved.

Against this background, we examined the effect of the above factor (2) and investigated how the existing ATR measures change before and after a term is published in a domain-specific dictionary. We admit that other factors might be more influential than dictionaries, but we leave the problem for further research. While there are studies on a general dictionary's impact on society (Read 1973 and Quirk 1973), there are few on the impact of domain-specific dictionaries on terminology, especially in the context of ATR.

Let us outline our survey method. The dictionaries and corpus of a certain domain are first prepared. The entry terms are regarded as terminology of that domain. Then, the ATR measures of the entry terms in the corpus are found for the cases of before and after dictionary publication. TFIDF, Nakagawa's FLR, C-value, MC-value, and simple document frequency are taken up as ATR measures. The domain in our investigation is women's studies because the first author is familiar with it.

2. Data

Below, we explain the dictionaries and corpus used for our investigations.

The dictionaries were two editions of the *Women's Studies Encyclopedia.* The first edition was published during 1989-1991 (because it consists of multiple volumes), and revised and expanded edition (henceforth called second edition) was published in 1999. The entries are regarded as women's studies terms in the present paper.[1] The first and second editions had 2,462 and 1,556 terms, respectively. They can be classified as follows:

- ◆ Terms which were in both editions (henceforth "MATCH"): 702 terms.
- ◆ Terms which were only in the first edition (henceforth "OLD"): 1,760 terms.
- ◆ Terms which were only in the second edition (henceforth "NEW"): 854 terms.

The temporal changes in ATR measures of these three were examined. In addition, we compared the following phrases in the corpus (which will be mentioned later) which were not listed in the dictionaries and whose POS patterns given by the Brill tagger were
{(Adjective|Noun)*Noun+} or
{(Adjective)*(Noun)+(Preposition)*(Noun)+}.

These phrases were regarded as ones that might be terms but are not listed in dictionaries (henceforth called MISMATCH). There were 19,561 mismatches.

The corpus was composed of abstract texts of the *Women's Studies International Forum* that were published between 1985 and 2003. The number of abstracts was 744, and each was composed of 143.9 words on average. They

[1] Persons' names were excluded.

were POS tagged by using the Brill tagger. To see the effect of publication in dictionaries on women's studies terms, we divided the abstracts into four equal parts around the publication years of the two editions (1989-91 and 1999). Henceforth, they are represented as P1, P2, P3 and P4, as shown in Figure 1. Table 1 shows the number of tokens in these parts of the corpus.

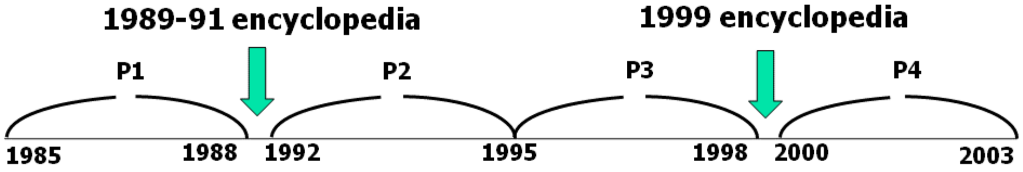


Figure1: Division of the corpus

Part	Year	Word Count
P1	1985-1988	31,472
P2	1992-1995	26,478
P3	1995-1998	29,592
P4	2000-2003	28,337

Table 1: Number of tokens in four parts of the corpus

3. Method

We examined the ATR measures of MATCH, OLD, NEW and MISMATCH in P1, P2, P3 and P4, respectively, and ascertained the temporal changes in their values. For instance, we examined how the TFIDFs of NEW terms changed from P3 to P4, i.e. how they changed after being listed as terms in the second edition of the dictionary.

As mentioned above, we dealt with five ATR measures. Their definitions are as follows:

1) TFIDF is defined as:

$$TFIDF(T) = f(T)\log\frac{D_0}{D(T)}$$

where *f(T)* is the number of occurrences of term candidate *T* in the corpus, *D(T)* is the number of documents where *T* appeared, and D_0 is the total number of documents of the corpus.

2) Nakagawa’s FLR (Nakagawa et al. 2003) is defined as:

$$FLR(T) = f'(T)(\prod_{i=i}^{L}(FL(t_i)+1)(FR(t_i)+1))^{\frac{1}{2L}}$$

where *f’(T)* is the number of times the term candidate *T* appeared in the corpus as an independent phrase or compound (in other words, it is not included in a longer phrase or compound), *L* is the number of words of *T*, t_i is the *i*-th constituent word of *T*, *FR*(t_i) is the total number of adjoining nouns on the left side of *t*, and *FR*(t_i) is the total number of adjoining nouns on the right side of *t*.

3) C-value (Frantzi et al, 2000) is defined as:

$$C\text{-}value(T) = \log_2|T|(f(T) - \frac{t(T)}{c(T)})$$

where |*T*| is the number of words of the term candidate *T*, *t(T)* is the frequency of occurrence of *T* in longer (already extracted as the above word formation) term candidates, and *c(T)* is the number of those candidate terms.

4) MC-value (Nakagawa et al, 2003) is defined as:

$$MC\text{-}value(T) = |T|(f(T) - \frac{t(T)}{c(T)})$$

where |*T*| is the number of words of the term candidate *T*, *t(T)* is the frequency of occurrence of *T* in longer (already extracted as the above word formation) term candidates, and *c(T)* is the number of those candidate terms. The MC-value is a modified version of the C-value (Frantzi and Ananiadou 1996; 1997) so that it can be applied to term candidates whose number of component words is one.

5) DF(*T*) is the number of documents where *T* appeared in the corpus.

4. Results and Discussion

4.1 Overall results

Table 2 shows the average values of the five ATR measures of four word types in four parts of the corpus. For instance, it shows that the average TFIDF of NEW terms dropped from 3.544 (in P3) to 2.893 (in P4); that means the average TFIDF of terms which are not listed in the first edition of the dictionary dropped after being listed in the second edition of the dictionary.

4.2 Impact of being listed in dictionaries

In Table 2, the average DFs of MISMATCH terms are apparently lower than those of other words. This means MISMATCH terms contain many rare terms which are quite different from MATCH, OLD, and NEW terms. To see how the dictionaries (or being listed in dictionaries) affect the ATR measures of terms, we analyzed the MISMATCH, MATCH, OLD and NEW terms for which C-values, MC-values, FLRs and DFs are in the range of 1 and 4 before being listed in the dictionaries.[2] By restricting the values of MISMATCH terms like this,

[2] As for TFIDF, because their values are relatively high, the range was set to 5 to 10.

many words that can actually be regarded as terms but are not listed in dictionaries would become contained in them. Therefore, by comparing their average TFIDF, C-value, MC-value, FLR and DF before and after being listed in dictionaries, we can partly see the impact of being listed in dictionaries. The results are shown in Table 3.

Table 3 indicates that the average TFIDF for NEW terms in P3 is 6.692 and that in P4 is 6.003. On the other hand, the average TFIDF of MISMATCH terms in P3 is 5.648 and that in P4 is 1.381. This means that while the TFIDFs of terms which were not listed in dictionary sharply declined, those of terms which were listed in the dictionary did not decline so sharply. The same can be said for the C-value, MC-value, FLR, and DF. The same can also be said for old and MISMATCH terms in P1 and P2. For instance, although the average DF of OLD terms dropped from 1.876 (in P1) to 1.019 (in P2), that of MISMATCH terms sharply dropped from 1.205 to 0.351.

Although these differences might be partly attributed to dictionary editors' foresight which words would enjoy longevity as terms and which would not, it seems reasonable to conclude that the dictionary has an impact on terms and that the effect is mostly to prevent their ATR measures from decreasing. In other words, the dictionary prevents the terms from losing their termhood that can be represented by ATR measures. Also, the fact that the values of ATR measures showed temporal changes indicates considering such changes might be effective for automatic term recognition. Especially, considering when the dictionaries were published and which terms they contained might be effective.

Words	Part	N	TFIDF	FLR	C-value	MC-Value	DF
MATCH	P1	702	7.444	4.181	0.094	2.449	1.623
	P2		6.349	2.786	0.067	1.988	1.358
	P3		7.173	3.779	0.062	2.313	1.588
	P4		6.125	3.176	0.043	2.039	1.342
OLD	P1	1,760	2.839	7.572	0.070	1.358	0.819
	P2		2.376	6.541	0.056	1.207	0.731
	P3		2.748	5.300	0.060	1.287	0.764
	P4		2.611	4.516	0.062	1.228	0.717
NEW	P1	854	3.151	1.489	0.018	1.006	0.811
	P2		2.865	1.202	0.011	0.881	0.660
	P3		3.544	1.773	0.025	1.112	0.799
	P4		2.893	1.032	0.018	0.849	0.680
MISMATCH	P1	19,561	2.196	1.497	0.170	0.786	0.474
	P2		1.910	1.295	0.150	0.688	0.415
	P3		2.165	1.547	0.169	0.777	0.461
	P4		2.059	1.407	0.160	0.744	0.440

Table 2: Average ATR measures of all terms

Words	Part	TFIDF		FLR		C-value		MC-value		DF	
		N	average	N	average	N	average	N	average	N	average
OLD	P1	118	6.505	88	1.899	34	1.904	149	2.202	209	1.876
	P2	118	3.447	88	0.870	34	0.814	149	1.317	209	1.019
MISMATCH	P1	3,994	5.614	2,371	2.031	1,598	1.316	3,936	2.410	4,527	1.205
	P2	3,994	1.238	2,371	0.353	1,598	0.087	3,936	0.363	4,527	0.351
NEW	P3	53	6.692	51	1.981	8	1.921	67	1.993	103	1.883
	P4	53	6.003	51	1.264	8	0.520	67	1.219	103	1.301
MISMATCH	P3	3,744	5.648	2,217	2.055	1,557	1.351	3,751	2.397	4,326	1.222
	P4	3,744	1.381	2,217	0.425	1,557	0.098	3,751	0.435	4,326	0.411

Table 3: Average ATR measures of terms whose DFs are 1-4 before being listed in dictionaries

4.3 Precision and Recall

We calculated the precision and recall of extracting MATCH, OLD and NEW terms from four parts of the corpus (P1, P2, P3 and P4) based on five ATR measures. The precision is defined as (the number of extracted terms which were listed in dictionaries)/(the number of extracted words). The recall is defined as (the number of extracted terms which were listed in dictionaries)/(the number of terms which were listed in dictionaries and existed in the corpus). The results concerning MATCH, OLD and NEW terms are shown in Figure 2, 3 and 4, respectively.

By comparing Figure 2 and Figures 3 and 4, we can see that the precision of extracting MATCH terms is generally higher than that of extracting OLD and NEW terms. Terms which are listed in many dictionaries are usually important terms and have higher termhood because (a) the fact that the words were listed in dictionaries influences the researchers and make them regard the words as their important terms, or (b) the words were important from the beginning and therefore they were listed in dictionaries. In that sense, MATCH terms are likely to have higher termhood than OLD and NEW terms both of which are listed in only one dictionary. The fact that the precision of extracting MATCH terms is generally higher than that of extracting OLD and NEW terms indicates that ATR measures are more effective for extracting terms which have higher termhood as they were originally designed for.

Figure 2 shows that the precision of extracting MATCH do not significantly differ among P1, P2, P3 and P4. By the above-mentioned reasons, MATCH terms can be regarded as the most important terms in women's studies. Such terms are stable during long period of time and their termhood might not be affected by dictionaries very much.

On the other hand, Figure 3 shows that the precision of extracting OLD increased from P2 to P4. Some of the OLD terms might be peripheral and less important compared to the MATCH terms and thus were not listed in the second edition of the dictionary.[3] These terms are easily influenced by being listed in dictionaries compared to the MATCH terms. The above-mentioned increase of precision can be partly attributed to this point, i.e., being listed in dictionary enhanced their termhood and it had gradually become easier for them to be extracted by ATR measures.

Among the ATR measures, FLR showed the largest precision increase concerning OLD terms. Because FLR assigns high score to the terms whose constituent units are used in many terms, it can be said that being listed in dictionary enhances the potential of word formation of each constituent unit of terms. In our case, many of the dictionary terms are single-word terms (i.e., terms and constituent units are equivalent). Therefore, being listed in dictionary is likely to enhance the potential of word formation of that term.

Incidentally, TFIDF showed the best performance in most cases while C-value showed the worst. It is because C-value is not designed for extracting single-word terms and many of the dictionary terms are single-word terms as we previously mentioned.

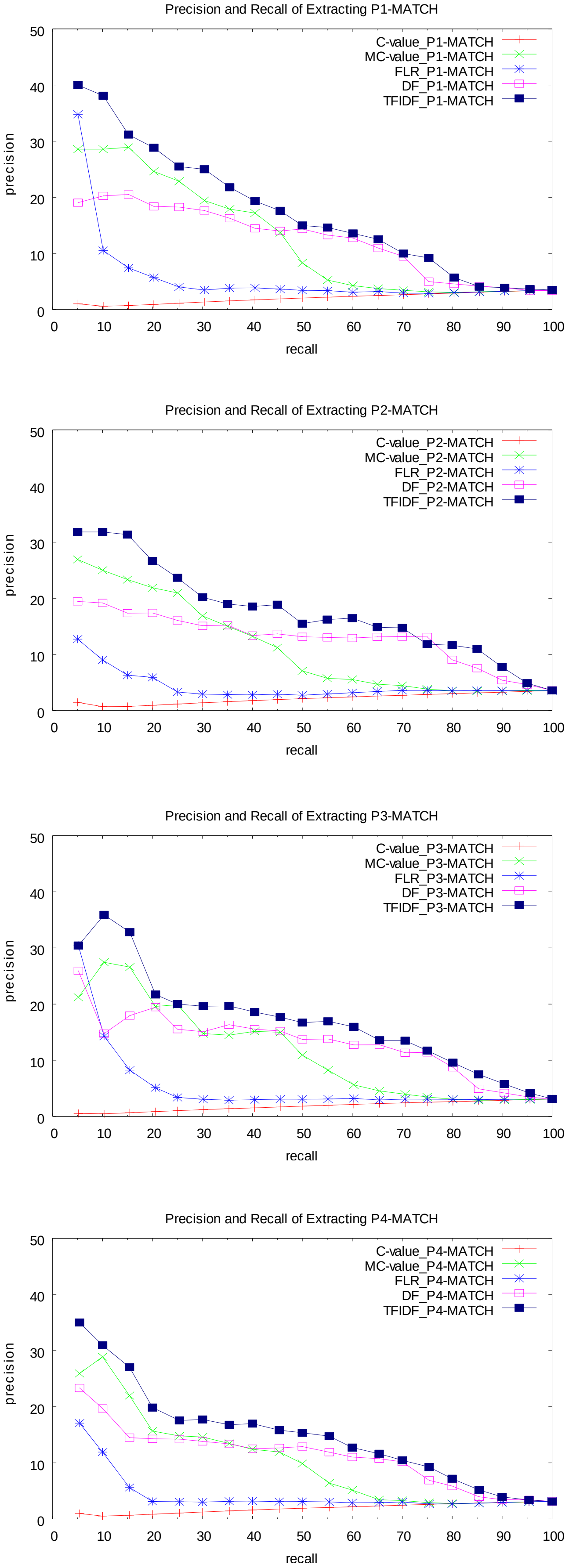
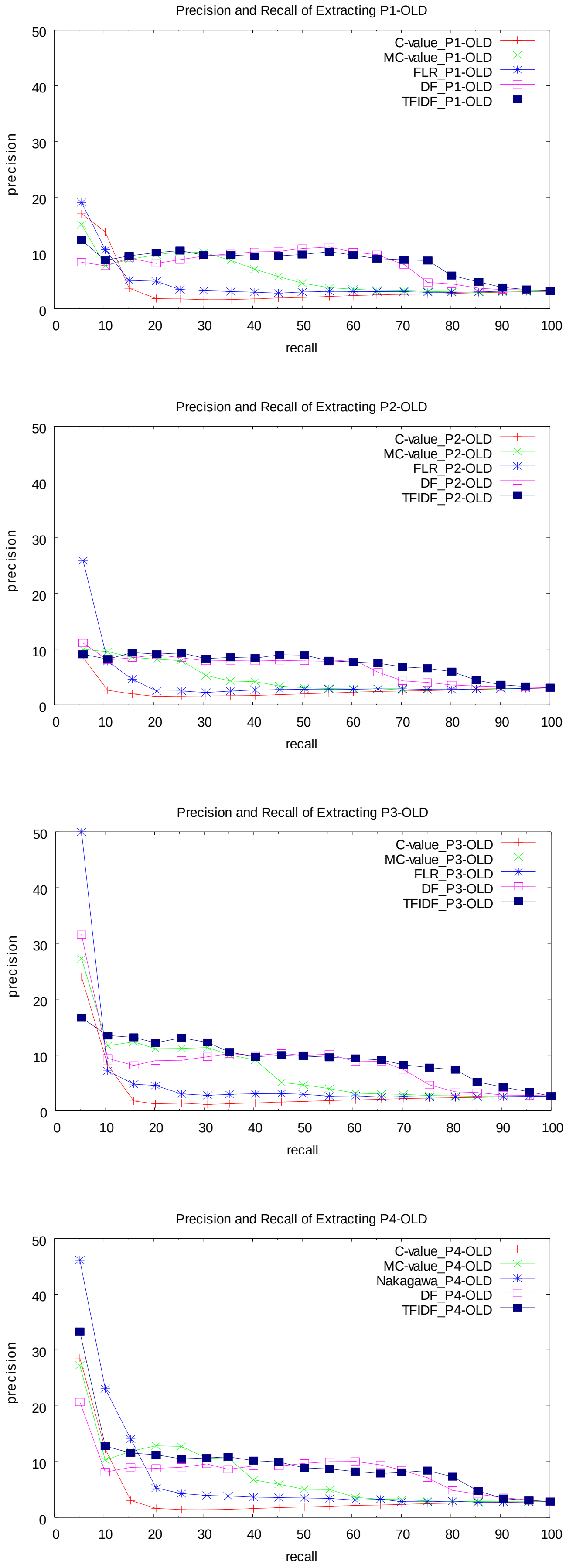


Figure2: Precision and Recall of Extracting MATCH

[3] We admit that some OLD terms became obsolete at the time when the second edition dictionary was published and thus were not listed in it.

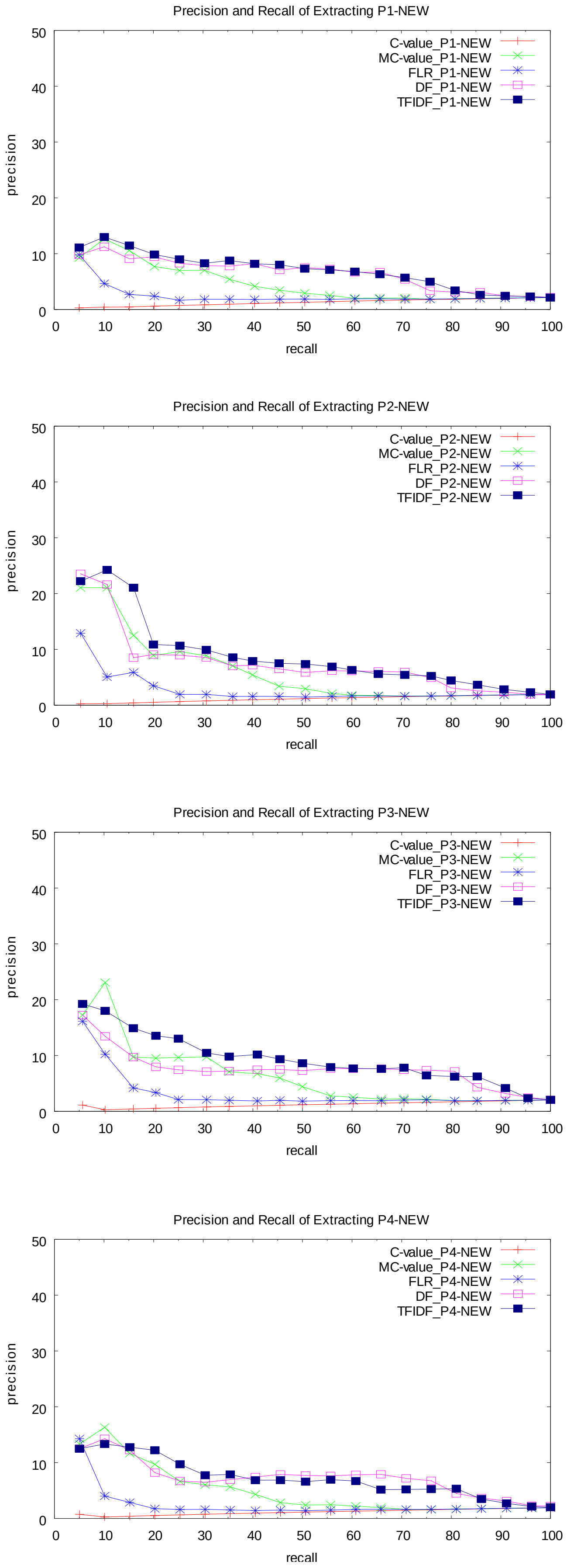

Precision and Recall of Extracting P1-OLD

Precision and Recall of Extracting P2-OLD

Precision and Recall of Extracting P3-OLD

Precision and Recall of Extracting P4-OLD

Figure3: Precision and Recall of Extracting OLD


Precision and Recall of Extracting P1-NEW

Precision and Recall of Extracting P2-NEW

Precision and Recall of Extracting P3-NEW

Precision and Recall of Extracting P4-NEW

Figure 4: Precision and Recall of Extracting NEW

4.4 Characteristic terms

The terms whose ATR measures sharply declined from P1 to P4 were related to (a) the women's movement, such as *black woman*, *liberation*, *racism*, *sexism*, *control*, *oppression* and *constraint*, (b) literature, such as *autobiography*, *diary*, *essay*, *myth*, *novel* and *poet*, and (c) legal terms.

The terms whose ATR measures sharply increased from P1 to P4 were related to (a) the problems facing women, such as *domestic violence* and *sexual harassment*, (b) life, such as *body*, *care* and *health*, (c) education, such as *teacher, training* and *higher education*, and (d) research, such as *case study, depth interview* and *qualitative study.*

5. Conclusions

To clarify the temporal aspect of terminology, we examined the dictionary's impact on terms using women's studies terms as data. The changes of averages of ATR measures before and after dictionary publication indicated that being listed in dictionaries gives some kind of longevity to terms and prevent them from losing termhood. From the changes of extraction precision and recall, it can be said that the peripheral or relatively less important terms are more likely to be influenced by dictionaries and their termhood increase after being listed in dictionaries. Among the termhood, the potential of word formation that can be measured by Nakagawa's FLR seemed to be influenced most.

We mentioned in Section 1 the influential factors on temporal aspect of terms other than dictionaries, i.e., journal papers, other media, academic societies, and linguistic factors. Their influence will be examined further.

References

Brill Tagger. Available: http://www.cs.jhu.edu/~brill/

Frantzi, K.T. and Ananiadou, S. (1996). Extracting Nested Collocations. In Proceedings of the l6th International Conference on Computational Linguistics (COLING-96), pp.41-46.

Frantzi, K.T. and Ananiadou, S. (1997). Automatic Term Recognition using Contextual Cues. In European Research Consortium for Informatics and Mathematics - Cross-Language Information Retrieval, pp.25-32.

Frantzi, K.T., Ananiadou, S. and Mima, H. (2000). Automatic Recognition of Multi-Word Terms: the C-value/NC-value method. International Journal on Digital Libraries, vol.3, no.2, pp.115-130.

Nakagawa, H., Mori, T. and Yumoto, H. (2003). Term Extraction Based on Occurrence and Concatenation Frequency. Journal of Natural Language Processing, vol.10, no.1, pp.27-45.

Quirk, R. (1973). The Social Impact of Dictionaries in the UK. Annals of the New York Academy of Sciences, vol.211, no.1, pp.76-88.

Read, A. L. (1973). The Social Impact of Dictionaries in the United States. Annals of the New York Academy of Sciences, vol.211, no.1, pp.69-75.

Tierney, H., ed. (1989). Women's studies encyclopedia. vol.1. Westport, Conn., Greenwood Press, 417p.

Tierney, H., ed. (1990). Women's studies encyclopedia. vol.2. Westport, Conn., Greenwood Press, 381p.

Tierney, H., ed. (1991). Women's studies encyclopedia. vol.3. Westport, Conn., Greenwood Press, 531p.

Tierney, H., ed. (1999). Women's studies encyclopedia. Rev. and expanded ed. Westport, Conn., Greenwood Press, 1,607p.