Relating Keywords to the 'Top Ten News of the Year' in Korean Newspapers

Jae-Woong Choe
Korea University / Seoul, Korea
jchoe@korea.ac.kr

Abstract

This paper takes an in-depth look at the relationship between mechanically extracted keywords and 'Top Ten News of the Year' compiled by the news editors. A previous study that briefly touched on the topic concludes there does not seem to exist any meaningful connection between the two. In this paper, we set up a more elaborate way of comparing and connecting the two, and argue that there is a certain reasonably good converging point. The corpus we make use of for our experiment is a subset of the Trend 21 corpus which is a collection of Korean major newspapers (2000-2013). For keyword extraction, log-likelihood ratio was made use of. Extraction of collocation for each keyword was needed, for which a version of Mutual Information was utilized. Finally a detailed comparison of the top ten news with the top 100 keywords was conducted from several points of view.

1 Introduction

There is a growing use of the keyword methodology as an analytic tool to efficiently analyze texts or corpora, and we can say it now has established itself as a viable, and, importantly, objective alternative to the traditional and rather introspective method of discourse or cultural interpretation (Scott & Tribble, 2006, Bondi & Scott, 2010; Archer, 2009; Baker et al., 2013). One issue that needs to be addressed is how the new methodology relates to the introspective way of selecting keywords from a given text or corpus. Are the two supposed to be different from each other? Then why so? Or do they have to be comparable or even identical to each other? And if they do, how can we test the comparability or convergence? In this paper we raise these questions on the basis of Korean newspaper corpora and some lists of 'Top Ten News (T10N for short)' compiled, presumably introspectively, by the newspaper editors at the end of every year. Specifically we compare the top 100 keywords with T10N, and see how well they converge with each other. Some previous studies seem to suggest a tentatively negative conclusion on the issue of any systematic relationship between the introspective and quantitative keywords in general (Bondi & Scott, 2010), or between quantitative keywords and T10N (Kim & Lee, 2011).

It should be noted at the outset that each item in T10N is not a keyword per se. Though it can trivially turned into a small set of keywords on the basis of the surface description of the news item, we take the item as an abstract concept to which a set of keywords can be mapped, thus making it possible to compare the quantitatively derived keywords and the more abstract key concepts that are selected introspectively. We argue that T10N provides an interesting testing ground for the significance of the keywords extracted or the role of the 'human factor' in the selection process of T10N.

This paper briefly reviews some of the previous studies on the issue at hand in Section 2, introduces the corpora used in Section 3, provides in Section 4 two lists of T10N to be analyzed, discusses methodological issues in Section 5, and reports the results in Section 6 with related discussion, which is then followed by conclusion.

2 Keywords: Introspection vs. Corpus

According to Stubbs (2010), there are three kinds of ‘keywords’, two of which are of our immediate concern in this paper: One is the ‘cultural keywords’, for example like the one compiled by Williams (1976/85), that are intended to capture the essence of a culture through selection of the terms or keywords that would represent some major aspects of the culture. It certainly would involve processes of understanding, interpretation, and abstraction on the part of the person that does the compiling job. In other words, they represent human interpretation and understanding of the phenomena, i.e., the culture. Let us call them ‘introspective keywords’, focusing on the methodology, so that they may cover not only the cultures but also other broad aspects of society. The other concept of keywords refers to the analytic tool mentioned in the previous section. They are called ‘statistical keywords’, which are extracted on the basis of the (relative) frequency distribution of words in the given corpora.

How are the two related to each other? Previous studies touched on the issue of the relationship between the two kinds of keywords, and suggested that they should be treated as separate kinds, little significant relationship being observed. For example, Stubbs (2010: 32) contends that they are “only loosely conceptually related, and perhaps only marginally compatible.” Scott (2010: 45) also states “[i]t is perfectly true that automatic analysis works differently from human identification in the case of keyness ….”

It may be true that human selection of the key words or phrases to capture the bigger picture of the society or culture is different from purely mechanical extraction of keywords from discourse. However it would also be true that the bigger picture is formed through discourse. It is mediated, communicated, and created through discourse. In other words, introspection based human keywords are not created from nothing; ideally they should reflect well the culture or society in question, and the discourse that constitutes and represents the society. Again ideally if we can get hold of the whole set of discourses which presumably reflect the culture and society as a whole, we can expect there would be a certain converging point between introspective keywords and quantitative keywords.

We will assume one of such case can be provided by news texts of a certain period of time. In particular, the top ten news summarize a year’s major events which had been reported in the news articles of the year. Suppose we have fairly large corpora that reflect the social events from which we can extract keywords on the one hand, and also human interpretation of the major events in the form of T10N on the other. How would they match up with each other then?

There is one previous study that dealt with such question. Kim & Lee (2011) compared T10N of 2009 and the keywords of the same year based on the Korean major newspapers, and found out only eleven out of 100 keywords match T10N. They concluded that “[the two] are more different from each other than would be expected (2011: 178). [Translation by the author]” While the presumably introspective selection of the major news of the year need not perfectly match the keywords, it is surprising that the major news does not seem to reflect the texts of the newspapers. This paper will take a careful look at this question again, using similar but more fine-tuned sets of data and different sets of tools for keyword extraction and interpretation. We find there is an interesting and meaningful converging point between the two, and even the items that do not match well seem to shed some light on how the human interpretation process works.

3 Target and Reference Corpora

T10N are selected annually, so it is reasonable to assume that the most relevant discourse would be the whole news of the year. For this experimental study, we are going to use the same set of T10N as that in Kim & Lee (2011), but we use different subsets as the target and reference corpora. In order to explain the difference, we first need to introduce the Trend 21 Corpus from which we take a subcorpus.

The Trend 21 Corpus is a collection of newspaper articles from 2000 to 2013 (Kim et al., 2011, Choe & Lee, 2014). All the newspaper articles in four major daily newspapers published in Korea (*Chosun, Dong-a, Joongang*, and *Hankyoreh*) constitute the corpus, and new data are processed and added after the end of each year when the data are provided by respective news media. The corpus currently contains over 600 million ‘ejels’, or chunks between spaces in Korean which typically consist of a content word and some agglutinating particles.

For the experiment in this paper we mainly use a subset of the corpus, primarily the data that cover the four year span (2006-2009) of a newspaper, *Chosun*, referring to a larger set when necessary. Specifically, we take the news texts of 2009 of *Chosun* as the target corpus, and those of 2006-8 of the same newspaper as the reference corpus. We assumed the corpus of the previous three years as the reference would be “moderate sized (Scott, 2010:52; See also Jeon & Choe (2009))”. This is different from Kim & Lee (2011) where they took all the news texts from the four major newspapers as the target and reference corpus, and then compared the results with T10N of *Chosun*. As is well known, newspapers may differ among themselves in terms of their respective stance on the social, cultural, and especially political issues (Baker et al., 2013). Thus assuming possibly different stances may influence the composition of each news texts and also the selection of T10N, we decided to compare the keywords from a particular newspaper with their own selection of T10N, thus limiting the effects of other factors to the minimum.

4 Lists of Top Ten News

There are typically two kinds of T10N compiled by each newspaper in Korea. One covers the national events, and the other international ones. *Chosun* had the following news items as their selection of T10N for the national and international major events of the year 2009, respectively.

Id	News item	Classification
cn1	Cardinal Kim and two former presidents passed away	Politics/ Death
cn2	Korea will host the G20 Summit in 2010	Foreign Affairs
cn3	Confrontation surrounding the Sejong City project and the four major river project	National Projects
cn4	The new North Korean leader Kim Jong Un, the second NK nuclear test	North Korea
cn5	The Naro space rocket launch failed	Science
cn6	Many large labor unions secede from *Minnochong* [the upper organization]	Labor
cn7	Media law passes the parliament	Media
cn8	Murderer Kang and the Nayoung case, Yongsan disaster not healed	Society
cn9	Golfer Yang wins over Tiger Woods, Kim Yu-na’ golden performance, Korean Soccer qualifies for the World Cup	Sports
cn10	The [Korean rice wine] makgeolli is all the craze everywhere	Life

Table 1: National Top Ten News (*Chosun*, 2009)

Id	News item	Classification
ci1	Expanding global economic crisis, weak dollars	Economic Crisis
ci2	Swine flu caused over 10,000 deaths	Epidemic
ci3	China's formidable economic growth	China
ci4	More American troops in Afghanistan	War
ci5	Lisbon Treaty, the EU's first president elected	Europe
ci6	Hatoyama assumes power in Japan, but the US-Japan relations seem murky	Japan
ci7	Copenhagen summit failed to meet the expectation	Environment
ci8	Pop emperor Michael Jackson dies	Entertainment
ci9	US women reporters detained in North Korea	US-NK relation
ci10	Tiger Woods’ infidelity scandal	Sports

Table 2: International Top Ten News (*Chosun*, 2009)

Note that the English translations of the news items provided in the above tables are not exactly the same as they appear in the newspaper but somewhat in an abbreviated and compact form, again to save the space. We also took the liberty of ignoring some metaphoric descriptions, and added some extra information so that those that are not familiar with the events may get a better grasp of them. For example, 'cn1' would be literally translated as "Kim Swu Hwan, Kim Dae Jung, Roh Moo Hyun … Major Figures in modern history are now in history," which simply refer to the deaths of the three major players in Korean politics and society for over 40 years. We also added the "Classification" column, again as a way to help the readers, especially those that are not familiar with the events described, to comprehend the overall picture as well as the characteristics of each event.

5 Methodology

5.1 Keyword extraction

The statistical procedures typically used for keyword extraction are Dunning's Log Likelihood (LL) and chi-square (Scott & Tribble, 2006). Some authors used T-score for the calculation (Kim & Lee, 2011). The standard text tools such as *WordSmith* (Scott, 2012) and *AntConc* (Anthony, 2011) provide keyword extraction procedures like log-likelihood and chi-squared, and it is generally known that there is not much difference between the two (Rayson, 2003; Bondie & Scott, 2010). In this paper, we used a version of LL described in Rayson (2003: 50). Let us suppose we have the following contingency table.

	Target	Reference
Frequency	a	b
Corpus Size	c	d

Table 3: Contingency table

Then the log-likelihood ratio is calculated as follows, where N refers to the total value of the four cells.

$$G^2 = 2\,(a\ln a + b\ln b + c\ln c + d\ln d + N\ln N - (a+b)\ln(a+b) - (a+c)\ln(a+\mathrm{c}) - (b+d)\ln(b+d) - (c+d)\ln(c+d))$$

The formula was implemented in a Perl script, rather than using any of the well-known tools, because the size of the data for the current study was rather huge and it was not easy, if not impossible, to handle them in the readily available tools. In order to confirm that the custom-made script works as expected, some test results were compared with those from *WordSmith* and *AntConc* on the basis of the same set of data, a Shakespearean play *Romeo and Juliet*, and there were minimal differences among the three results.

Since our concern in this paper is the topic rather than the style of the data, we limited our search to the words/morphemes that are nouns (/NNG) and proper names (/NNP), ignoring all the other categories.

5.2 Collocation extraction

For many of the extracted keywords, it was obvious which T10N item each of them belong to. But for many others, the connection was not that clear. There were several reasons for this. For one thing, a keyword may be ambiguous. For example, 지원[jiwon] in Korean may either mean 'support' or 'application', and we need to figure out in which sense the word was selected as a keyword. Another reason is that it was difficult to decide in which context a certain keyword was used. 중소기업[jungsogieop] means 'small and medium sized enterprises', and it is difficult to know whether it has anything to do with T10N or not. A third reason is that some keywords may be linked to more than one item in the T10N list. 오바마(Obama), as President of the most influential country in the world, can obviously be related to many news items. Finally there were a few keywords with baffling identities. 김정운[Kim Jung-un], apparently a personal name, was listed as a keyword, and it was not clear at first why the name cropped up as a keyword.

These problems can be solved if we take the context into consideration, of course. A widely used method is to browse the keywords in the KWIC style. But when there are so many data to be checked, a more efficient method is called for which will succinctly summarize the contextual information. One such method would be collocation, which looked good enough for our purpose so we made use of it in this study. Thus for each keyword, a set of collocation words, or more exactly a set of morphemes were gathered that co-occur in the same news item.

There are well-known collocation extraction methods like the t-test and Mutual Information. While the t-test seems to have some issues with low frequency words, Mutual Information has been considered too skewed to them, assigning a very high value, for example, to a bigram whose members occur only once in the given corpus. In this paper, we used a version of Mutual Information, called Log-Frequency based Mutual Dependency (LFMD, Thanopoulos et al., 2002), which is designed to add some frequency effect to Mutual Information. The metric is given below, which was again implemented in Perl:

$$D_{LF}(w_1w_2) = D(w_1,w_2) + \log_2 P(w_1w_2)$$

where D is:

$$D(w_1, w_2) = I(w_1, w_2) - I(w_1 w_2) =$$
$$= \log_2 \frac{P^2(w_1 w_2)}{P(w_1) \cdot P(w_2)}$$

6 Results and Discussion

Once the keywords were extracted and sorted in a descending order of their LL value, we checked each of the top 100 keywords for possible matchup with the twenty items of the national and international T10N provided in Tables 1 and 2. The collocation word list for each keyword was constantly consulted in the process. A sample of the table used for the process is provided in Appendix at the end of this paper.

6.1 Keywords that relate to the national T10N

30 out of the top 100 keywords were found to be linked to the national top 10 news. Their mutual relationship is provided in the following table, where the T10N news items and their related keywords are shown side by side. Each keyword is followed by its English gloss, and then its rank in the 100 list shown in parenthesis.

Cat: Cla.	keywords(rank)
cn1: Politics/ Death	서거/Death(22), 조문단/Condolence_delegation(74), 조문/Condolence(76), 분향소/Memorial altar(81), 국민장/National_funeral(93)
cn3: National Projects	세종시/*Sejong*_City(1), 충청/*Chungcheong*(16), 원안/First_draft(40), 대강/Major_rivers(49), 해양부/Maritime(68), 사업/Business(72), 국토/Country(99)
cn4: North Korea	오바마/*Obama*(6), 보즈워스/Bosworth(38), 김정운/Kim_Jong_Un(87), 도발/Provocation(92), 버락/*Barack*(95)
cn5: Science	나로호/*Naroho*(18), 관제/Control(23), 발사체/Projectile(55)
cn5/cn4	로켓/Rocket(15), 발사/Launch(32)
cn6: Labor	노조/Labor_union(26), 노총/Trade_unions(34), 민노총/*Minnochong*(44), 탈퇴/Withdrawal(82)
cn7: Media	미디어/Media(65)
cn8: Society	강호순/*Kang_Hosun*(66)
cn9: Sports	김연아/*Kim_Yuna*(80)
cn10: Life	막걸리/Rice_wine(19)

Table 4: National T10N and their matching keywords

It seems like each item in T10N is reasonably well represented in the top 100 keyword list. Each item, except for ‘cn2’, has at least one keyword that supports its selection. Half of the national T10N (‘cn1’, ‘cn3’, ‘cn4’, ‘cn5’, ‘cn6’), or five out of top 6 news, are linked to at least three top 100 keywords. Top three major news given in Table 4, namely ‘cn1’, ‘cn3’, and ‘cn4’, have at least five matching keywords. Overall, we might be able to say that the top five items in Table 4 support rather strongly the convergence between the introspection based major news and the statistically derived keywords.

The bottom four items in Table 4 is not that well supported by the list of top 100 keywords, but still each finds a keyword in the list that can be linked. We will come back to the missing one 'cn2' in Section 6.3.

The first three names that appear as keywords for 'cn4: The new North Korean leader Kim Jong Un, the second NK nuclear test' show why collocation information is needed for proper classification. 오바마(Obama) would rather be expected to be linked to some international news, and no doubt Obama, as President of the most influential country of the world, would be featured in many international news. However, when the collocation words of the name were checked, a crucial one seems to be '핵[haek]/nuclear'. Obviously, as much as 'Obama' appeared in many other news, the name was significantly associated with the word 'nuclear' and the most noteworthy mention of the word 'nuclear' in Korea in 2009 was in the context of the North Korean nuclear test. The same applies to another name "(Steven) Bosworth" in 'cn4'. The words that collocate with it are such as '방북[bangbuk]/visiting North Korea', '회담[hoedam]/talks', '대북[daebuk]/to North Korea', and '특사[teuksa]/special envoy', clearly revealing his role as a special US envoy handling the NK nuclear issue. Finally, '김정운[Kim Jung-un]', apparently a personal name, was listed as a keyword, and even to a person that is well versed in Korean national affairs the name looked puzzling at first. Its collocation revealed the name refers to the newly emerging North Korean leader. His name was initially wrongly identified as 김정운 in the media, rather than the correct 김정은 as was later to be known through the North Korean media, befitting to the secrecy and mystery that surrounds the country.

Even with some collocation information, there were truly ambiguous cases, and thus we had to add an extra classification category 'cn5/cn4' in Table 4. The two keywords associated with the category, namely '로켓[rokes]/Rocket(15)' and '발사[balsa]/Launch(32)', when their collocated words were considered, were clearly linked either to the failed launching of the spacecraft *Naroho* in the South, and to the launching of the missile *Daephodong* in North Korea. We therefore tentatively classified it as belonging to 'cn5/cn4'.

6.2 Keywords that relate to the international T10N

The same number of keywords, namely, 30 out of the top 100 keywords was found out to be linked to the international T10N, as shown below.

Cat: Cla.	keywords(rank)
ci1: Economic Crisis	위기/Crisis(9), 회복/Recovery(10), 불황/Recession(20), 글로벌/Global(25), 금융/Finance(61), 부양책/Stimulus_package(67), 회복세/Recovery(75), 침체/Downturn(79), 신흥국/Emerging_country(91), 조정/Adjustment(96)
ci2: Epidemic	신종플루/Swine_flu(2), 플루/Flu(4), 신종/New_type(8), 백신/Vaccines(14), 접종/Vaccination(36), 확진/Confirmed(43), 타미플루/*Tamiflu*(45), 인플루엔자/Influenza(62), 감염/Infection(64), 독감/Flu(73), 바이러스/Virus(86), 의료/Medical_care(94), 환자/Patients(97)
ci6: Japan	하토야마/*Hatoyama*(11)
ci7: Environment	녹색/Green(3), 저탄소/Low-carbon(46), 기후/Climate(50), 코펜하겐/*Copenhagen*(57), 온실/Greenhouse(78), 친환경/Eco-friendly(85)

Table 5: International T10N and their matching keywords

Almost all of the 30 keywords were linked only to the three international T10N items. The items 'ci1: Expanding global economic crisis, weak dollars' and 'ci2: Swine flu caused over 10,000 deaths' were the two prominent international news of the year that were amply reflected in the keywords. They were national news as well as international ones as Korea was also affected by both the economic crisis and the epidemic, and thus people in Korea were keenly following the news.

Economic crisis and the swine flu epidemic are the two pieces of news that affect the lives of the general public very much, and obviously the news media were clearly aware of them, thus dealing with them very widely and repeatedly as the related keywords show. Likewise, global warming and the subsequent climate change is one of the grave issues that largely bother the minds of the general public. Thus the 2009 United Nations Climate Change Conference, or the Copenhagen Summit was apparently covered well in the newspaper as the keywords in 'ci7: Copenhagen summit failed to meet the expectation' show in Table 5.

The other news item in the table is about Japan. Hatoyama became the first Prime Minister from the modern Democratic Party of Japan in 2009, defeating the long-governing Liberal Democratic Party. The power change in Japan, a closely related neighboring country to Korea, was an obviously newsworthy item to Koreans, and so was covered accordingly in the news media.

6.3 T10N that do not have any matching top 100 keywords

Among the national T10N news items 'cn2' was the only exception that did not have any supportive words in the top 100 keywords. The description of 'cn2' is 'Korea will host the G20 Summit in 2010' as shown in Table 1. The key phrase in the description is 'G20', but it turns out that the tagger wrongly analyzed it as 'G' and '20'.

G20 이 G/SL+20/SN+이/JKS

Since the source of the problem was located, it was possible to get the log-likelihood value for the expression "G20" separately. Had it been treated as a single unit, its LL value would be 2287.92, which means "G20" would rank as the ninth item in the top 100 keyword list (See Appendix). Every national T10N in Table 1 is supported by at least one associated keyword.

Item	O1	1%	O2	2%	LL
Word	1277	0.01	267	0.00 +	2287.92

Table 6: Log-likelihood value for 'G20'

On the other hand, as for the international T10N news items, only four of them could find their linked keywords in the top 100 keyword list, as Table 5 shows. Note that three of them, the most heavily covered ones ('ci1', 'ci2', 'ci7') in the media, concern global issues which would also affect the lives of the local general public. It is highly likely they would have made national T10N even if they were not covered by the other list.

Among the rest of the news items in Table 2, five of them deal with regional issues like China, US-led war in Afganistan, Europe, Japan, and a US-North Korean issue. The other two concern well-known popular figures like Michael Jackson the pop star and Tiger Woods the sports star. Only one of these seven items has a single related word in the top 100 keyword list. It is obvious that not particularly many news articles were written on the global scale topics in the newspaper and yet the editors felt they should be included in T10N. We will come back to this point later.

6.4 Top 100 keywords that do not belong to any of the T10N items

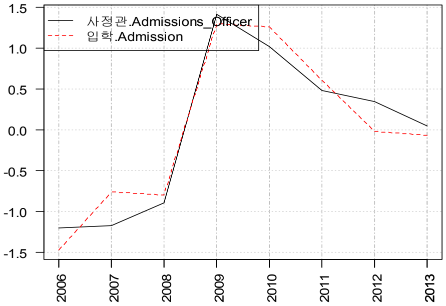
There were 40 keywords in the top 100 list that were left out of the national and international T10N. Four of them were included due to some other factors than the news stories themselves. For example, the use of the word '편집자[pyeonjipja]/ Editor' seems to have spiked up in 2009, but it was exclusively used as part of the editorial comments to some of the articles, rather than as part of the news stories. Many of the other keywords were used individually, having little to do with other words in the top 100 keyword list. However, there were several clusters of keywords each of which seemed to point to a particular event or topic.

cat	keywords(rank)
o_edu	사정관/Admissions_Officer(12), 교과부/MOE(28), 전형/Exams(35), 사교육/Private_tutoring(37), 입학/Admission(41), 수능/SAT(58), 성적/Grades(71), 지원/Application(77), 모집/Recruitment(90)

o_job	잡월드/Job_World(5), 취업/Job_finding(39), 일자리/Jobs(47), 비정규직/Non-regular_workers(60), 중소기업/Small_business(70)
o_housing	보금자리/Bogeumjari_housing(27), 수도권/Metropolitan(30)
o_IT	스마트폰/Smartphone(63), 트위터/Twitter(84)

Table 7: Keyword clusters each of which points to a particular topic

The keywords in the 'o_edu' category concerns college entrance system, particularly the newly introduced one by universities in Korea that seemed to be gaining huge momentum, urged by the Ministry of Education. Education, especially the college entrance system is everybody's concern in Korea, and even a slight change in the system has huge repercussions on the society in general. Obviously, the new "Admissions Officer" system was one of the top national issues in 2009 and thus was much talked about in the media. The following chart shows that the use of the keywords '사정관[sajeonggwan]/ Admissions_Officer' and '입학[iphak]/ Admissions' greatly increased simultaneously in 2009 in the four major newspapers in Korea.



Another hot topic which is everybody's concern is unemployment or difficulty of getting a job. Growing number of the unemployed has become a social issue. It was a much talked about issue of the year again, as the keywords in the 'o_job' in Table 7 show. Incidentally, the first keyword of the category, '잡월드[jabwoldeu]/ Job_World(5)' turned out to be the name of the website created by the particular newspaper, *Chosun*, together with other institutions, as part of a social campaign to help the unemployed to find a job. Out of 1,161 occurrences of the word in the four major newspapers in 2009, the vast majority (1,151) have appeared in *Chosun*. The other two categories in Table 7, along with their keywords, again have a lot to do with everyday life of the ordinary people: a new housing project in the metropolitan area, and newly introduced popular IT items like smartphones and the twitter.

The rest of the keywords, 14 of them, seemed to deal with individual issues separately, and did not aggregate well among themselves. One thing to note before we close this section is the personal or pen names that showed unusual degree of keyness though not related to any of the T10N.

cat	keywords(rank)
p_invst	박연차/*Bakyeoncha*(13), 건호/*Geonho*(98)
p_ent	장자연/*Chang_Jayon*(48)
p_soc	미네르바/*Minerva*(100)

Table 8: Keywords of proper names

The category 'p_invst' is related to the political scandal that implicated a former president, which many believe eventually led to his suicide. The two keywords in the category refer to the principal figures in the scandal, close associates of the late president. The other two categories in Table 8 are again related some social and political scandal.

Many of the keyword clusters or keyword discussed in this section could have made the national T10N list but the editors chose otherwise.

7 Conclusion

So how well do the statistically derived keywords and the introspection based T10N converge? Based on the results of our analyses, over 60 percent of the top 100 keywords make positive contribution to the convergence. Seen from the opposite point of view, national T10N is well supported by the keywords while international T10N is markedly less so. So we can conclude that though the two do not match with each other perfectly, they converge reasonably well.

Then what is the source of the difference between them? Here we provide some speculative remarks. One thing that influences the introspection based selection or decision of T10N is a higher abstraction process involved. For example, 'cn3: National construction projects' is an abstraction over more than one separate event or project, and so are 'cn1: Politics/Death' and 'cn4: North Korea'. Secondly, the introspection based selection is likely to be influenced by the historical context. The choice of 'death' as the top national news of the year in 'cn1', rather than the second or the third or even below, would make more sense if we take it into consideration that the three people involved have left a huge impact in recent history of Korea. So it is not just their death, but in a sense the end of era in Korean political and social history that mattered in the selection.

The third factor that apparently plays a role in the selection of the T10N is the geographical balance, especially in the case of the international T10N, namely from 'ci3' to 'ci6' in Table 2, for which there were not to be found any related keywords in the top 100 list. The fourth factor is the sectional or topical balance of the newspaper (Tables 1, 2). Otherwise it is rather difficult to explain the inclusion of the last four items ('cn7' to 'cn10') in Table 4 at the expense of other events that are more prominently reflected in the linked keywords. The final factor that seems to matter is what we might call topic subsumption. Although the name '박연차/*Bakyeoncha*' and '건호/*Geonho*' appeared particularly frequently in 2009 (See Section 6.4), the event was eclipsed by a much bigger related news which was the death of the former president allegedly involved.

References

Laurence Anthony. 2011. AntConc (Version 3.2.4) [Computer Software]. Tokyo, Japan: Waseda University.

Dawn Archer. Ed. 2009. What's in a Word-list?: Investigating Word Frequency and Keyword Extraction, Ashgate.

Paul Baker, Costas Gabrielatos & Tony McEnery. 2013. Discourse Analysis and Media Attitudes: the Eepresentation of Islam in the British Press. Cambridge University Press.

Marina Bondi & Mike Scott. Ed. 2010. Keyness in Texts. John Benjamins Publishing Company.

Jae-Woong Choe, Do-Gil Lee. 2014. Trends 21 Corpus: Public Web Resources and Search Tools. Studies in Korean Culture 64. pp. 1-20. [In Korean]

Jieun Jeon & Jae-Woong Choe. 2009. A Key word Analysis of English Intensifying Adverbs in Male and Female Speech in ICE-GB. Proceedings of the 23rd PACLIC. City University of Hong Kong. pp. 210-219.

Heunggyu Kim, Beom-mo Kang, Do-Gil Lee, Eugene Chung, Ilhwan Kim. 2011. Trends 21 Corpus: A Large Annotated Korean Newspaper Corpus for Linguistic and Cultural Studies, Digital Humanities 2011, June 19-22, 2011. Stanford University.

Ilhwan Kim and Do-Gil Lee. 2011. Automatic Keyword Extraction and Analysis from the Large Scale Newspaper Corpus Based on t-score. Korean Linguistics 53. pp. 145-194. [In Korean]

Paul Rayson. 2003. Matrix: A Statistical Method and Software Tool for Linguistic Analysis through Corpus Comparison. Ph.D. thesis, Lancaster University.

Mike Scott. 2010. Problems in Investigating Keyness, or Clearing the Undergrowth and Marking out Trails…. In Bondi & Scott, 2010. pp. 43-58.

Mike Scott. 2012. WordSmith Tools version 6. Lexical Analysis Software.

Mike Scott & Christopher Tribble. 2006. Textual Patterns: Key Words and Corpus Analysis in Language Education. John Benjamins Publishing Company.

Michael Stubbs. 2010. Three Concepts of Keywords. In Bondi & Scott, 2010. pp. 21-42.

Aristomenis Thanopoulos, Nikos Fakotakis, George Kokkinakis. 2002. Comparative Evaluation of Collocation Extraction Metrics. Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas, Spain.

Raymond Williams. 1976/85. Keywords: A Vocabulary of Culture and Society. Fontana Press.

Appendix: A sample of the table used for the linking process

No	Morph/Eng	cat	cl	freq	LL	mi terms
1	세종시/Sejong_City	NNP	n3	2718	6289.892	원안:6.351 수정:5.622 총리:3.741 충청권:3.645 정부:3.521 행정:3.436 정:3.283 정운찬:3.17 도시:3.044 문제:2.965 수정안:2.776
2	신종플루/Swine_flu	NNG	i2	2298	5958.051	감염:4.611 환자:4.036 확진:3.994 타미플루:3.732 백신:3.471 접종:3.07 예방:2.739 독감:2.437 확산:2.242 바이러스:2.023 인플루엔자:1.938
3	녹색/Green	NNG	i7	3360	3475.75	성장:6.96 저탄소:6.219 에너지:5.153 산업:3.913 환경:3.681 친환경:3.259 사업:3.239 계획:3.15 기술:3.049 정부:2.999 추진:2.819
4	플루/Flu	NNG	i2	1185	3072.164	신종:9.552 감염:4.83 환자:3.805 인플루엔자:2.745 바이러스:2.706 감염자:2.608 백신:2.374 확진:2.073 타미플루:1.424 질병:1.199 개학:0.984
5	잡월드/Job_World	NNP	o_job	1151	2984.009	취업:6.114 채용:5.953 기업은행:5.949 중소기업:5.947 구직자:4.945 청년:3.659 인재:3.225 사이트:2.88 일자리:2.662 조선일보:2.381 이력서:2.083
6	오바마/Obama	NNP	n4	6809	2764.495	대통령:8.779 미국:8.015 행정부:7.894 미:7.856 버락:7.26 Obama:6.738 백악관:6.165 부시:5.716 회담:5.163 클린턴:5.017 핵:4.839
7	자전거/Bicycles	NNG	o_trans	5081	2548.662	도로:5.841 이용:2.606 구간:1.664 공원:1.64 설치:1.627 시민:1.553 계획:1.387 조성:1.376 한강:1.248 전용:1.12 교통:1.034
8	신종/New_type	NNG	i2	1552	2536.461	플루:9.552 감염:5.071 인플루엔자:5.043 환자:3.894 바이러스:3.47 백신:2.806 감염자:2.531 확진:2.155 독감:2.038 질병:1.899 대유행:1.68
9	위기/Crisis	NNG	i1	11859	2173.943	경제:9.897 금융:9.757 글로벌:8.092 미국:7.913 세계:7.78 말:7.735 시장:7.583 정부:7.377 이후:7.192 상황:7.119 기업:6.979
10	회복/Recovery	NNG	i1	5174	2040.508	경기:7.848 경제:7.081 금융:6.56 위기:6.445 상승:6.118 시장:6.095 말:6 전망:5.992 이후:5.834 투자:5.7 침체:5.663
11	하토야마/Hatoyama	NNP	i6	859	2002.735	유키오:5.55 총리:5.512 일본:3.652 자민당:3.501 일:3.383 정권:2.57 민주당:2.305 오자와:1.95 오카다:1.942 후텐마:1.661 오키나와:1.661
12	사정관/Admissions_Officer	NNG	o_edu	1163	1856.23	입학:8.134 전형:6.388 관제:5.97 선발:4.816 사정:4.811 면접:4.604 학생:4.215 서류:3.985 입시:3.969 대학:3.727 성적:3.423
13	박연차/Bakyeoncha	NNP	p_invst	1164	1800.486	검찰:6.197 태광실업:6.008 수사:5.891 회장:5.15 노무현:5.119 대검:4.722 박:4.707 노:4.608 게이트:4.467 중수부:4.13 소환:4.036
14	백신/Vaccines	NNG	i2	1460	1766.166	접종:6.082 독감:4.666 바이러스:4.212 예방:3.842 신종플루:3.471 감염:3.367 인플루엔자:3.062 신종:2.806 타미플루:2.737 플루:2.374 녹십자:2.337
15	로켓/Rocket	NNG	n5/n4	1431	1662.941	발사:7.967 우주:4.796 장거리:4.348 발사체:4.244 위성:4.191 미사일:4.188 나로호:4.001 북한:3.625 러시아:2.686 단:2.59 인공위성:2.587