Special issue: Leading ICT technologies in the Information Explosion

Technical Note

Computing the potential lexical productivity of head elements in nominal compounds using the textual corpus

Kyo KAGEURA
Graduate School of Education, University of Tokyo

ABSTRACT

In this paper, we propose a model/method for observing the potential lexical productivity of head elements in nominal compounds, and compare the productivity of a few high-frequency head elements in a technical corpus. Much work has been done on various aspects of nominal compounds. Most of this work, however, has been devoted to "syntagmatic" aspects at various levels, such as semantic compositionality, possible variations, lexical cohesiveness, etc. while the "paradigmatic" aspects of nominal compounds have received relatively little attention. By providing a "paradigmatic" perspective for analysing nominal compounds, this paper aims to help build a truly integrative approach to analysing and processing nominal compounds.

KEYWORDS

lexical productivity, paradigmatic perspective, head elements, nominal compounds, terminology

1 Introduction

This paper aims to provide a concept and method for observing the potential lexical productivity of head elements in nominal compounds based on textual corpora. Many studies have been carried out on various aspects of nominal compounds. Most of them have been devoted to "syntagmatic[1)]" aspects at various levels, such as structural and/or semantic analysis [1]–[4], variations [5]–[7], and lexical cohesiveness [8] [9]. On the other hand, little work, if any, has been done on the lexical productivity of nominal compounds, such as the productivity of head elements.

Though this situation may reflect the fact that most NLP-related "applications" are concerned with texts or discourse and may therefore be considered reasonable, it is also true that an actual manifestation of a text or a discourse always presupposes what could have been manifested in its place. Also, intuition[2)] tells us that there is at least some correlation between the potential lexical productivity of an element and the semantic nature of its compounds, such as their transparency or compositionality. For instance, it might be that the more productive the element is, the more transparent the semantic relation between the element and its company tends to be. Suppose, for instance, the nominal compounds "A X" and "B X" (where "A" and "B" are modifiers and "X" is the head) are seemingly non-compositional semantically. It would be revealing, if not critical, to know whether the head "X" is extremely productive or not in analysing the nature of the non-compositionality of the compounds "A X" and "B X". By addressing the issue of potential productivity, we shift the target of theoretical observation from the textual sphere or the use of compounds to the lexicological sphere [11].

As such, the potential productivity of the nominal compounds of an element should provide a basic frame of reference for analysing individual cases of nominal compounds in which the element is used. Though the work reported here falls short of fully exploiting the potential for the integrated processing of nominal compounds (partly due to the fact that potential lexical productivity has been virtually unexplored so far)[3)], it of-

Received August 25, 2008; Revised October 27, 2008; Accepted October 28, 2008.

kyo@p.u-tokyo.ac.jp, http://panflute.p.u-tokyo.ac.jp/~kyo/

1) We use this word in the broadest possible sense.

2) Kripke argued that "some philosophers think that something's having intuitive content is very inconclusive evidence in favor of it. I think it is very heavy evidence in favor of anything, myself. I really don't know, in a way, what more conclusive evidence one can have about anything, ultimately speaking" [10]. We agree with his opinion, and resort to our intuition in interpreting potential productivities of various head elements.

DOI: 10.2201/NiiPi.2009.6.6

3) In fact, in studies of automatic weighting of compound terms, the use of information related to lexical productivity as well as its possible integration

fers useful insight into approaching the phenomenon of nominal compounds.

2 Potential lexical productivity

2.1 The concept of lexical productivity

Let us define here the "potential lexical productivity" of the head element of compound nouns[4]. Take, for instance, two lexical elements w_1 and w_2, and define two corpus-based quantitative measures as follows:

$f(i,N)$: the token frequency of an element w_i occurring as a head of nominal compounds in a corpus of size N, i.e. the token frequency of nominal compounds the head of which is w_i.

$d(i,N)$: the type number of w_i occurring as a head of nominal compounds in a corpus of size N, i.e. the number of types of compounds the head of which is w_i. It is sometimes called the "extent of use" [18].

Suppose now that $f(i,N)$ and $d(i,N)$ for w_1 and w_2 in a corpus of size N are as follows:

	$f(i,N)$	$d(i,N)$
w_i	1,000,000	50,000
w_j	10,000	1,000

In this case, we might say that w_i is more lexically productive simply because $d(i,N)$ for w_i is higher, which means w_i makes more nominal compounds than w_j. However, if $f(i,N)$ and $d(i,N)$ for w_1 and w_2 in a corpus of size N are as follows,

	$f(i,N)$	$d(i,N)$
w_i	1,000,000	1,000
w_j	10,000	1,000

we might very well say that w_j is *more* lexically productive because it produces 1,000 nominal compounds while occurring only 10,000 times in texts, while on the other hand w_i only produces 1,000 types of nominal compounds although it occurs a million times in the corpus; the latter should be evaluated as having poor productivity.

This intuition in turn raises doubts concerning the conclusion in the first case, i.e. that w_i is more lexically productive as it produces 100 times more nominal compounds than w_j. An argument can be made against this claim as follows: if w_j had occurred in the corpus with the same frequency as w_i, we do not know which of the two would have produced more nominal compounds.

This argument shows that, although $d(i,N)$ can be regarded as representing the lexical productivity of the head element, $f(i,N)$ also affects the observation of lexical productivity: if other conditions are equal, an element tends to be regarded as less productive if it has a higher $f(i,N)$ value. This derives from the simple intuition that (1) the more frequent the token occurs, the greater the type number tends to become, and (2) if we talk about the *lexical* productivity of the head elements of nominal compounds, we should not be affected by the token occurrence of elements in textual corpora, because this is determined by how dominant the elements are in discourse and has little to do with *lexical* productivity *per se*. It is this intuitive notion of lexical productivity that we would like to formalise here.

Let us also note that when we talk about "if w_j had occurred ..." we are introducing the *potential* productivity as opposed to a snapshot of the productivity in the corpus.

We can thus define the *potential lexical productivity* of a head element as:

> The potential ability of a head element to construct nominal compounds when it occurs in a text, as distinct from the characteristic of how frequently the element can be used in a text.

This implies that lexical formation is detached from the actual token use of the element[5], while it still depends on its use in texts[6].

According to this definition, the potential lexical productivities of the two head elements w_i and w_j could be compared by normalising the token frequency of the element. To remove the effect of token occurrence $f(i,N)$ from $d(i,N)$, using the type-token ratio seems, at first glance, to be a convenient method:

$$\frac{d(i,N)}{f(i,N)}$$

However, it is widely recognised that this and other types of simple measure depend systematically on the sample size [14]. Thus it can only be used, at best, to produce a rough snapshot of lexical productivity.

One way to remove the factor of corpus size completely is to estimate the number of nominal compounds when the head element occurs an infinite number of times in texts, i.e.:

$$d(i) = d(i,\lambda N), \quad \lambda \to \infty \tag{1}$$

This fits the definition of potential lexical productivity, in which the factor of token occurrence of the head element in texts is completely removed.

with syntagmatic information has been discussed [12]. However, one can see that this has only been done heuristically; this is partly because the concept of lexical productivity has not been properly explored.

[4] The definition itself is not necessarily confined to heads.

[5] Work in psycholinguistics has shown that the subjective frequency of a simplex word depends on the type number of compounds it constructs [13].

[6] The formal correspondence of the notion of the dependency of lexical productivity on texts will be introduced in the next section.

Alternatively, we can observe the developmental profiles of $d(i, \lambda N)$ for changing λ. Compared to $d(i)$, observation of the developmental profiles has the advantage of allowing the nature of productivity to be analysed more exploratively. For the aim of comparing the productivities of the different elements w_i and w_j, for instance, it is necessary to compare $d(i, X)$ and $d(j, Y)$ where $f(i, X) = f(j, Y)$. Below, we will use the developmental profiles of $d(i, \lambda N)$ to observe the potential lexical productivity, instead of using $d(i)$. Given the lack of understanding of the basic nature of potential lexical productivity, we believe this is a sensible choice.

2.2 Models and methods for measuring potential productivity

In order to define $d(i)$ and $d(i, \lambda N)$ properly, we need to take into account the distributions of and around the head element more consistently. In view of the concept of potential lexical productivity, the relevant distributions for defining $d(i)$ and $d(i, \lambda N)$ are:

(1) The distribution which gives the probability of a head element in the sphere of potential lexical productivity of head elements, which is given by the ratio of the number of different compounds that contain w_i to the total number of all the potential compounds taking any heads. This distribution is determined by $d(x)$ for all the head elements x in the potential lexicon.

(2) The distribution which gives the probability of observing a particular compound which contains w_i, when w_i occurs in a sample (this is defined for each focal element)[7].

(3) The distribution which gives the probability of observing w_i in the document set.

Letting p_{w_i} be the occurrence probability of w_i in texts, and C_i be the sample space $\{i_1, i_2, i_3, ..., i_{d(i)}\}$ of the distribution of compounds that contain w_i with probability $p_{(c)_{i_k}}$ given to each compound i_k, and assuming the combination of binomial distribution, we have:

$$\mathrm{E}[f(i, N)] = Np_{w_i}, \tag{2}$$

$$\mathrm{E}[d(i, N)] = \sum_{m=1}^{Np_{w_i}} \sum_{k=1}^{d(i)} \binom{Np_{w_i}}{m} p_{i_k}^{m} (1 - p_{i_k})^{Np_{w_i} - m}. \tag{3}$$

Note that Np_{w_i} here may not necessarily be a natural number, but we can practically regard it as being a natural number, because the occurrence of linguistic items is always discrete. What is given in the data is the empirical value for $d(i, N)$, with the empirical distributions of what actually occurs among C_i in the document set.

[7] This corresponds to the dependency of lexical production on the actual use of the head element in texts, as mentioned above.

Thus the task to be solved now is to estimate $d(i, \lambda N)$ for arbitrary λ. For $\lambda < 1$, we can empirically obtain the developmental profiles of $d(i, \lambda N)$ by means of random subsampling of the original data. However, it would be more useful if we could observe the data size range of $\lambda > 1$ as we are talking about *potential* lexical productivity. In order to do so, we can use binomial interpolation and extrapolation [15].

Binomial interpolation and extrapolation provides an estimate of the number of compounds that take w_i as a head, i.e. $E[d(i, \lambda N)]$ as follows:

$$\mathrm{E}[d(i, \lambda N)] = \mathrm{E}[d(i, N)] - \sum_{k=0}^{\infty} (-1)^k (\lambda - 1)^k \mathrm{E}[d_k(i, N)] \tag{4}$$

where $d_k(i, N)$ indicates the number of compound words that take the head w_i and which occur k times in a corpus of size N (i.e. the original corpus). Thus, if we allow ourselves to estimate $\mathrm{E}[d_k(i, N)]$ and $\mathrm{E}[d(i, N)]$ by

$$\hat{\mathrm{E}}[d_k(i, N)] = d_k(i, N) \tag{5}$$

and

$$\hat{\mathrm{E}}[d(i, N)] = d(i, N) \tag{6}$$

respectively, it is possible to calculate the number of nominal compounds $d(i, \lambda N)$ that take w_i as a head for $\lambda > 1$ as well.

3 Observing potential lexical productivity

3.1 Data and setup

In order to observe the potential lexical productivity of head elements, we used 1,025 articles published in the *Journal of the American Society for Information Science (and Technology) (JASIST)* over a period of 17 years (1986 to 2002). Though 17 years might be long enough to affect the lexical productivity of some head elements, we assumed that this corpus gives a synchronic slice of language representing the field of information science.

The data was tagged using the Brill tagger [16], and the two word sequences consisting of NN, NNS, NNP were extracted as nominal compounds (candidates)[8]. We then cleaned up the data. The token and type number of nominal compounds thus obtained are shown in Table 1. In accordance with the shift of the theoretical target of observation from the textual sphere to the

[8] We have here omitted compounds with more than three constituent elements as they are relatively few in number and the direct modification of the heads can in many cases be obtained by compounds with two constituent elements.

Table 1 Token and type number of nominal compounds in *JASIST*.

	token	type
number	567,821	227,067

Table 2 Token and type number of nominal compounds with the 10 most frequent head elements.

	$f(i,N)$	$d(i,N)$	d/f
system	13072	1422	0.108
retrieval	7136	454	0.064
term	5493	725	0.132
model	3845	764	0.199
process	3799	571	0.150
science	3721	194	0.052
information	3461	1356	0.391
analysis	3436	591	0.172
search	3283	695	0.212
data	2799	885	0.316

lexicological sphere, we regard this corpus as a representative text range through which the unique and singular lexicological sphere can be addressed. As we will see later in relation to the treatment of possible statistical errors, this treatment limits the interpretative framework but at the same time theoretically anchors the observations to the lexicological sphere, which is consolidated as a singular historical event.

Among the data thus extracted, we selected the 10 most frequent head elements for observing the potential lexical productivity. Table 2 gives $f(i,N)$ and $d(i,N)$ as well as the type-token ratio of these 10 elements. It is possible to classify these head elements in terms of the status of the concepts they represent in the field of information science, as follows:

Base concepts of the domain: "information" and "data" (and perhaps "system"[9]) represent very basic concepts in the field of information science;

Specialised core concepts of the domain: "retrieval", "term" and "search" represent more specific but essential concepts in the field;

General concepts: "model", "analysis", "process" and "science" represent more general concepts[10].

[9] "system" is a difficult case because depending on the context it can be regarded as a base concept, a specialised core concept or even a general concept with respect to the field of information science.

[10] The nature of "science" seems to be slightly different from the other three concepts.

We can observe from the type-token ratio that within the data some elements such as "information" and "data" are highly productive while such elements as "retrieval" and "science" are not productive.

3.2 Basic developmental profiles

For these 10 most frequent head elements, we applied binomial interpolation and extrapolation. Figure 1 shows the developmental profiles of the 10 elements (the x-axis is taken by f, thus some elements can be observed only for a partial range of other elements), up to twice the original data size.

The developmental profiles show that "information", then "data", are *consistently* most productive, and "science", then "retrieval", are *consistently* least productive. The developmental profiles of the other six elements are concentrated around the profile of "system". Upon closer inspection of the trajectories, some interesting tendencies can be observed:

(1) Comparing "search" and "model", whose profiles are very similar, the curve of "model" flattens out more quickly than that of "search"; we can observe that the d("*model*", f) becomes smaller than d("*search*", f) at one point.

(2) Comparing "analysis" and "process" on the one hand and "term" on the other, the curves of "analysis" and "process" flatten out more quickly than that of "term"; d("*process*", f) becomes smaller than d("*term*", f) within the range of observation, while d("*analysis*", f) might well become smaller than d("*term*", f) if we extrapolate further.

Thus we may be able to make the generalisation that elements representing specialised core concepts tend to retain their productivity more than elements representing general concepts, when the range of observed productivities are similar. This, however, is not saying much given that (a) "model" is — and is expected to be at least according to a visual extrapolation — still much more productive than "term" and "retrieval", and (b) "analysis" and "process" are — and are expected to be — still far more productive than "retrieval".

Incidentally, this observation would ideally be accompanied with a discussion of possible errors and variances that necessarily accompany statistical estimations. However, from the viewpoint of lexicology, these variances and errors should be dealt with as a sampling issue of textual corpora that represent the lexicological sphere and not as sampling issues of lexical items of a given textual corpus. Though this is a choice that depends on the singularity of the lexicological sphere and the theoretical status of the study of lexicology [17], and thus can be argued against from the statistical point of view, we nevertheless take this standpoint in this paper and make clarification of this issue as the future task

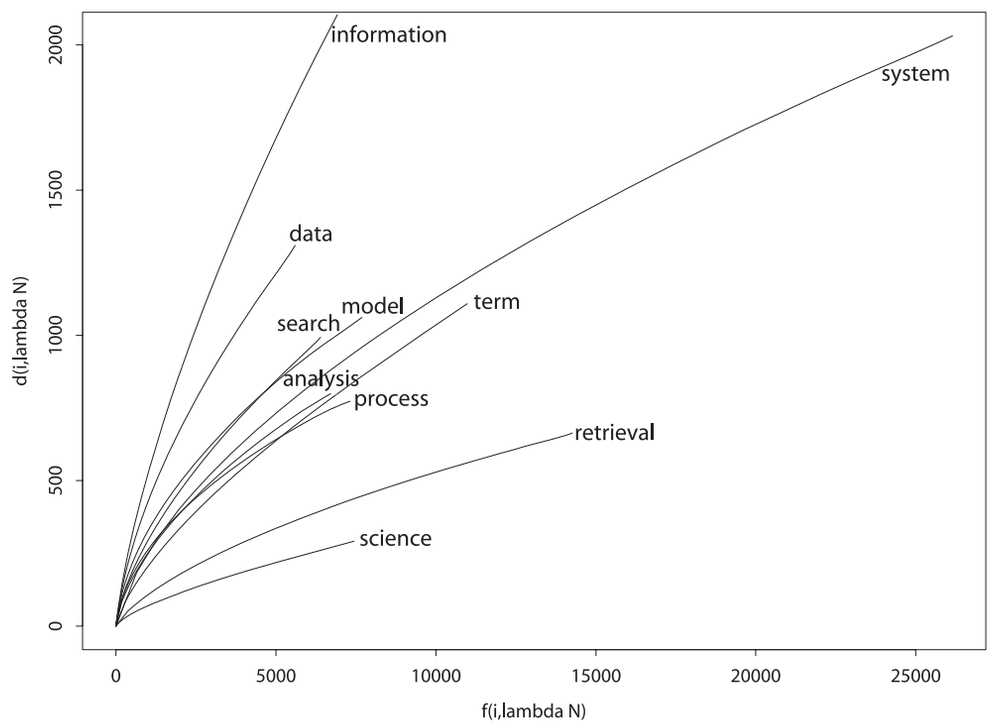

Fig. 1 Developmental profiles of the 10 most frequent head elements.

to be dealt with at the level of sampling of a textual range, rather than subsampling of a given textual corpus. These two are, from the viewpoint of lexicology, essentially different and the issue of choosing textual range may well be qualitative rather than straightforwardly statistical.

3.3 Removing highly lexicalised compounds

These head elements, as long as they are so frequently used in the field of information science, may well produce some nominal compounds which are highly lexicalised, to the extent that these compounds are regarded as functionally equal to simple words[11]. We can hypothesise this with respect to the lexical productivity as follows:

> Conspicuously "lexicalised" compounds do not affect the productivity of the head element of these compounds.

For instance, "natural language processing" is so constantly used and the concept it represents so fixed, it can be seen as highly lexicalised. We may therefore be able to claim that this compound does not affect the production of nominal compounds which take the head "processing".

We thus checked the distribution of nominal compounds for each of the 10 head elements, and removed some compounds from the data according to the following criteria: (a) the token occurrence of the compound is more than two percent of all the compound tokens

Table 3 Highly lexicalised compounds for the 10 head elements.

head	modifiers of compounds
system	retrieval, information, IR, expert
retrieval	information, document, text, image
term	search, index, query, indexing, subject
model	retrieval, space, data, IR, boolean
process	search, retrieval
science	information, computer, library
information	—
analysis	citation, content
search	web, subject, information, keyword, boolean
data	citation

[11] For instance, the fact that "information retrieval" can be abbreviated as "IR" shows that "information retrieval" is such a compound.

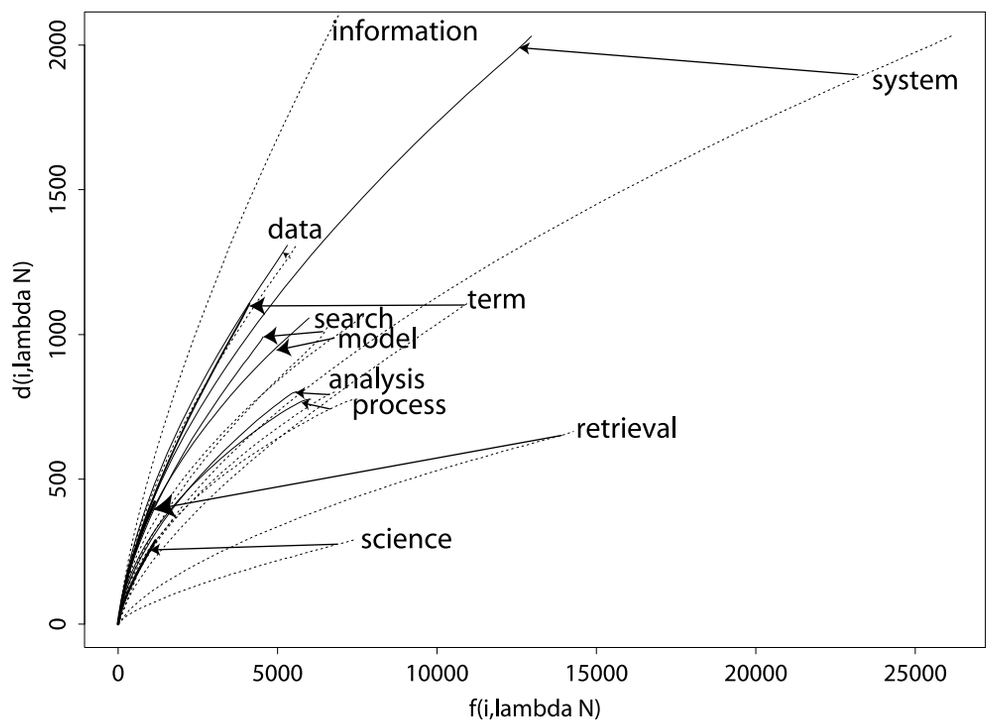

Fig. 2 Developmental profiles of the 10 most frequent head elements, before and after the highly lexicalised compounds are removed.

that take the same head[12], and (b) they represent fixed concepts of the domain (determined by referring to a handbook and two dictionaries in the field of information and library science). Table 3 shows the compounds thus removed for each of the 10 head elements. These compounds are assumed to be lexicalised and do not affect the productivity of compounds of the head elements.

Figure 2 shows the changes in the developmental profiles of $d(i, \lambda N)$ for the 10 elements when these highly lexicalised compounds are removed. It can be noted that, after the removal of these compounds, the productivities of "system", "retrieval" and "information" (and "science") becomes much greater (as indicated by the arrows in Figure 2). On the other hand, the changes in the developmental profiles of the productivities of "search", "model", "analysis", and "process" are rather small. In the case of "data" and "information" the changes are minimal (or zero).

Figure 3 underlines that, after the removal of highly lexicalised compounds, the lexical productivities of base and specialised core concepts of the domain show rather similar developmental profiles; they show not only high productivities but also a general tendency not to drop off. General concepts, on the other hand, also show similar developmental profiles; they show low productivities and the developmental profiles are inclined to flatten out. "model" is a somewhat gray case, but still the developmental curve seems to start dropping off only toward the end of the observed range, compared to such elements as "search" or "system"[13].

3.4 Some nature of head elements

Taking into account the lexical productivities shown in the developmental profiles as well as the change in the developmental profiles of productivities when highly lexicalised elements are removed, it is possible to define a few groups of head elements in which the nature of productivity correlates with the status of the concepts they represent:

(1) Elements which are very productive in general but does not produce highly lexicalised compounds. Among the 10 head elements observed here, "information" and "data" belong to this class. The head elements of this class can be said to represent base concepts of the domain.

[12] Because we tend to observe a gap in grouped frequency distributions at around this point.

[13] $d(i)$ estimated using LNRE models [18] shows that the order of the potential number of compounds with "model" as a head is much smaller than that of "search", though we cannot claim much reliability in relation to this point.

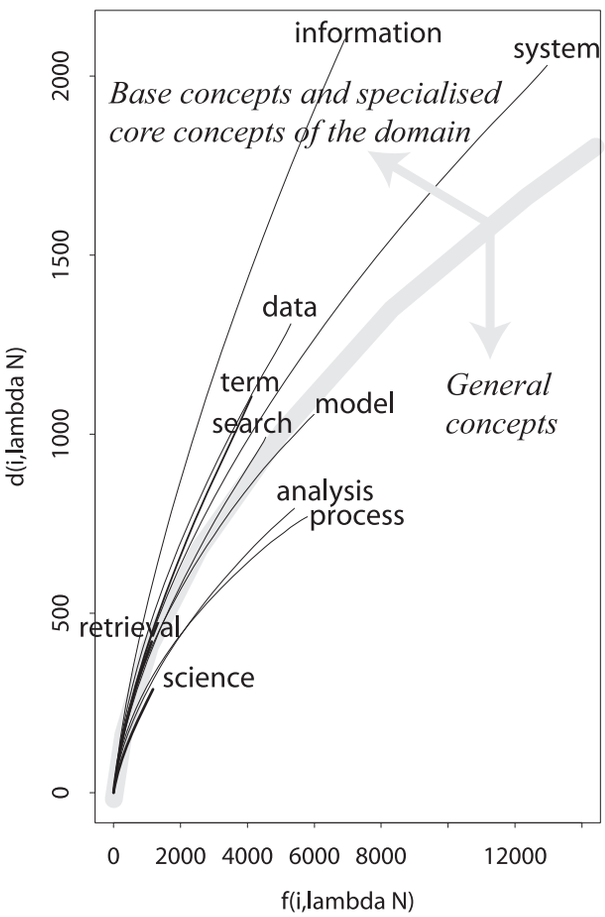

Fig. 3 Developmental profiles of the 10 most frequent head elements., after the highly lexicalised compounds are removed.

(2) Elements which are very productive in general and produce some highly lexicalised compounds that are used very frequently in discourse. "retrieval", "term" and "system" belong to this class. This class represents specialised core concepts of the domain. In this regard, "search" is a gray case, which does not appear to have produced conspicuously lexicalised compounds[14]. This class of head elements is expected to be the basic "pool" for the production of domain-dependent lexicalised compounds.

(3) Elements which are less productive than those classified under (1) and (2), and do not produce highly lexicalised compounds. The elements falling into this class are "model", "analysis" and "process". This class represents general concepts, important for representing various operations or aspects of the field.

(4) Elements which are less productive than those classified under (1) and (2), but do produce a few highly lexicalised compounds. Among the 10 elements observed here, "science" belongs to this class. This class represents general concepts.

These claims are dependent on two mutually supporting assumptions/hypotheses, i.e. (a) the status of the concepts of the 10 elements in the field of information science can be roughly divided into three types as listed in 3.1 and (b) highly lexicalised compounds can be determined by their high frequency of occurrence in texts and the stability of the concepts they represent. As these two assumptions/hypotheses are not so well established — though they fit our intuition — we have to admit that these claims are not conclusive. Nevertheless, observation of these 10 elements shows that these claims are highly plausible and makes a strong case for their further exploration.

4 Conclusions

This paper introduced the concept of potential lexical productivity as well as a method for empirically observing productivity. On the basis of this concept and method, we have shown that the observation of lexical productivity is very useful in characterising the nature of head elements.

The pursuit of consolidating the lexicological sphere, of which the present work is a part, is of great importance both theoretically and practically. Theoretically, as Saussure argued, lexicon is the core constituent of *la langue* [19], and computational linguistic work has become matured enough so that it is time this issue was fully explored. Practically, with the realisation of global accessibility to virtually unlimited textual resources on the Web, the lexical application is shifting from trying to extract as much as possible from textual corpora to construct coherent set of lexical items that fits to applications. Put it differently, what is currently requested is not the extraction of lexical items, but the construction of coherent dictionaries [20].

Though what we have clarified in this paper still falls short of incorporating potential lexical productivity in the processing and treatment of nominal compounds, it should ultimately prove useful as even the partial and heuristic use of lexical productivity information in term weighting was reported to be highly promising [12]. As a direct extension of the research reported here, we will address the following issues in the next step:

(1) The choice of textual corpora as data through which the lexicological sphere is observed. This issue involves the treatment of errors and variances in the observations explored in this paper using binomial extrapolation;

[14] The fact that the term "search" in the field of information science is used to represent a looser and broader concept than "retrieval" may explain this.

(2) The qualitative examination of the assumptions we adopted in this paper, e.g. in relation to the lexicalisation and the setting of the threshold in section 3;
(3) The extension of the head elements to be observed. We observed only 10 head elements in this paper. In order to draw more generalised and stronger conclusions, a larger number of head elements should be observed under the same assumptions.

As mentioned at the end of 3.4, the conclusions drawn in this paper in relation to the nature of the head elements depend deeply on a few interrelated assumptions, which should themselves need separate confirmation. As such, what is reported in this paper is in its very embryonic stage. However, given the dearth of work on the lexicological sphere, we believe that this study is useful as a stepping stone from which we can explore the nature of constituent elements of compounds within the lexicological sphere.

Acknowledgement

The research reported here was partially supported by the 2008 NII research cooperation project "Research on the modelling and reconstruction of lexicographical space from textual corpus". I would like to thank Professor Akiko Aizawa for discussion over the lexicological aspects of complex terms.

References

[1] C. Fabre, "Interpretation of nominal compounds: Combining domain-independent and domain-specific information," *COLING-96*, pp.364–369, 1996.

[2] M. Johnston and F. Busa, "The compositional interpretation of nominal compounds," E. Viegas, Ed., *Breadth and Depth of Semantic Lexicons*. Dordrecht: Kluwer, 1998.

[3] M. Lauer, *Designing Statistical Language Learners: Experiments on Noun Compounds*. Ph.D thesis, Macquarie University, 1995.

[4] K. Takeuchi, et. al. "An LCS-based approach for analyzing Japanese compound nouns with deverbal heads," *Computerm 2002*, pp.64–70, 2002.

[5] B. Daille, et. al. "Empirical observation of term variations and principles for their description," *Terminology*, vol.3, no.2, pp.197–257, 1996.

[6] C. Jacquemin, *Spotting and Discovering Terms through NLP*. Cambridge, Mass: MIT Press, 2001.

[7] K. Kageura, et. al. "Parallel bilingual paraphrase rules for noun compounds: Concepts and rules for exploring web language resources," *Proceedings of the Fourth Workshop on Asian Language Resources*, pp.54–61, 2004.

[8] K. Church and P. Hanks, "Word association norms, mutual information and lexicography," *Computational Linguistics*, vol.16, no.1, pp.22–29, 1990.

[9] F. Smadja, "From n-grams to collocations: An evaluation of Xtract," *ACL-91*, pp.279–284, 1991.

[10] S. A. Kripke, *Naming and Necessity*. Cambridge, Mass: Harvard University Press, 1972.

[11] K. Kageura, "Quantitative portraits of lexical elements," *Proceedings of the Third International Workshop on Computational Terminology*, pp.75–78, 2004.

[12] H. Nakagawa, "Automatic term recognition based on statistics of compound nouns," *Terminology*, vol.6, no.2, pp.195–210, 2000.

[13] H. Baayen, et. al. "The morphological complexity of simplex nouns," *Linguistics*, 35, pp.861–877, 1997.

[14] F. Tweedie and H. Baayen, "How variable may a constant be?" *Computers and the Humanities*, vol.32, no.5, pp.323–352, 1998.

[15] I. J. Good and G. H. Toulmin, "The number of new species, and the increase in population coverage, when a sample is increased," *Biometrika*, vol.43, no.1, pp.45–63, 1956.

[16] E. Brill, http://cs.jhu.edu/~brill, 1997.

[17] K. Kageura, "The dynamics of phenomena and the dynamics of data — on the relationship between events and structures in terminology —," *Mathematical Linguistics*, vol.22, no.7, pp.281–302, 2000.

[18] H. Baayen, *Word Frequency Distributions*. Dordrecht: Kluwer, 2001.

[19] Saussure, F. de *3ème Cours de Linguistique Générale*. Geneve: Bibliothèque Publique et Universitaire, 1910–1911.

[20] K. Kageura, "Terminological lexicons and terms in context: The translator's perspective," Dieng-Kuntz, R. and Enguehard, C. eds. *7e conference Terminologie et Intelligence Artificielle*. Grenoble: Presses universitaires de Grenoble, pp.1–10, 2007.

Kyo KAGEURA

Kyo KAGEURA, PhD, is working as an associate professor of Library and Information Science Course, Graduate School of Education, the University of Tokyo. His main research interests are: formal modelling of media and language structure, terminology and library-related applications of natural language processing. He is currently leading the Shiitake project, in which a translation aid system for online volunteer translators is developed.