Some Creative Aspects of Nominalization: An Analysis of Hapax Legomena in English

Junya MORITA*

1. Introduction

It has been suggested by various researchers, including Clark and Clark (1979) and Aronoff (1980), that the words that are coined on the spur of the moment should be differentiated from those stored in the lexicon and the productive types of words are chiefly created depending on the context in which they occur. Central to this phenomenon is context-dependent online nominalization, which is very active and plays some significant syntactic and pragmatic roles. It is also pointed out that hapax legomena provide a reliable and objective indicator of what words are not stored in long-term memory (Baayen and Renouf (1996)). The aim of the present study is to elucidate certain creative aspects of nominalization by analyzing the hapax nominals extracted from a large-scale corpus. The outline of this article is as follows: after pointing out the significance of hapax legomena and the method of their research in §2, we will investigate in §3 the internal and external structures of the hapax nominals attested and explicate how these structures are related to the pragmatic functions—discourse cohesion, focus, and brevity. Section 4 will concentrate on analyzing contextually-motivated hapax nominals, and show how each type of the nominal forms is to be selected in a relevant situation to carry out the function of discourse cohesion.

2. Hapax Legomena and the Method of Their Research

2.1. Significance of Hapax Legomena

It has been recognized by the Standard and GB theorists that lexical items consist of simple and complex words, each of which is inserted into the relevant node of a phrase structure to make a D-structure (cf. Allen (1978: 197)). Importantly, regular complex words are not stored in the lexicon, but they can be produced on demand by word formation devices; for example, the novel words *co-disposal* and *claw-removal* are aptly coined in a particular context. Jackendoff (1997: 131-133) specifically states that inventive type of word formation like this gets involved in working memory, where items are essentially composed by free combinatory rules. The present study will be concerned exclusively with creative online word formation, since it is this type of word coinage that crucially contributes to constructing the simplified and elegant lexicon.

To obtain a proof of such a creative facet of word formation, it is vital to reveal precisely what items are not lexically listed. A number of psycholinguistic experiments have shown that highly frequent

complex words are characteristically recovered from the mental lexicon faster and more accurately than low frequency ones (Hay (2003: 77-81)). This entails that while complex words with high frequency are permanently stored in the mental lexicon and retrieved without being accessed via word formation rules, complex words with very low frequencies of occurrence are generally composed by the rules and accordingly their extraction without error would require more time. As a criterion of low frequency word, hapax legomena play a most useful role. A hapax legomenon is a word which appears only once in a sample and is hence coined once for a particular occasion by a single speaker. This extremely low frequency word is crucially used in measuring "productivity"―the probability of encountering new formations―on the assumption that a complex word of very low frequency is generally created by a word formation rule (Baayen and Renouf (1996)). For this reason, examination of hapax nominals is indispensable for investigating the true nature of creative nominalization.[1]

2.2. Method of Research and the Result

In order to collect as many nominal hapaxes as possible, we have looked for the hapaxes in the British National Corpus (BNC). By repeatedly using the "wild card" function of a research engine, the frequency of complex words ending with nominal suffixes has been checked to find nominals of token frequency 1.[2] Nominalization is here defined as the process of forming a nominal on the basis of a verb phrase, and so a gerund is left out from our discussion, which correlates with a clause that can encode aspectual distinctions. Our focus is also on the nominal which denotes a process or activity.

Further, the nominal form of [self-V_{suf}]$_N$ (e.g. *self-notification*), the nominal whose base verb is of zero frequency (e.g. *disidentify* (token frequency 0) ― *disidentification* (hapax)), or the nominal with a minor spelling (e.g. *despatchment*) is not the subject of our investigation. The reason for the exclusion of *self*-compounds should be noted here. A reflexive pronoun is used both as argument and intensifier, as in *he takes himself too seriously* and *he must do the work himself,* and in the argument use it has to refer back to the subject of the same clause. As will be shown below, the connection of the first unit of a compound to the expression of the same referent outside the compound is an important factor for online nominalization, but a synthetic compound which takes *self* as argument is peculiar in that the first element *self* cannot relate to an expression outside the compound or the NP containing it. Besides, there is a fair number of hapaxes of this sort; 41 hapaxes of this kind have been detected in the BNC and all except one incorporate argument-type of *self*. If *self*-type and ordinary type of compounds were put together, it would make the nature of hapax nominals very unclear. *Self*-compounds are worth examining, but they should be considered separately from usual synthetic compounds.[3]

As a result of the research, as many as 248 types of nominal hapaxes are identified in the 100-milion-word corpus. It should be noted that of the 248 hapaxes discerned, 170 have no entry in the *The Oxford English Dictionary* (2nd ed.). As discussed in the last section, a complex word of very low frequency is generally coined by some word formation device. Even though a hapax nominal is listed in the comprehensive dictionary, it can be newly reorganized on the spot in the speaker's mental lexicon, especially in the case where the base verb is transparent; the meaning of the verb is determined by the

combination of the meanings of its components (e.g. *overgeneralization* and *prettification*). Hence the BNC hapax nominals are highly likely to be coined on demand even if a part of them has an entry in *OED*.[4] Further, since words which are not registered in *OED* can be judged as novel words, the BNC hapax nominals which have no entry in the large dictionary are very highly likely to be generated online by word formation devices, that is to say, they are true online nominalizations. That a significant portion of the hapax nominals listed in *OED* can be recomposed and a considerable number of the hapaxes extracted are not registered in the comprehensive dictionary strongly reinforces the view that hapaxes are principally constructed online in working memory.

3. The Internal and External Structures of Nominalizations and Their Functions

3.1. Internal and External Structures

In this section, the detected hapaxes will be classified in terms of syntactic patterns and each class will be examined in detail. Internally, nominalization can be divided into five main types, as depicted in Table 1: (I) (Det) + derived N (e.g. *(the) destruction*), (II) (Det) + derived N + of + NP (*(the) destruction of the city*), (III) (Det) + $[N + V_{suf}]_N$ (*(the) city-destruction*), (IV) possessive + derived N + of + NP (*Caesar's destruction of the city*), (V) possessive + $[N + V_{suf}]_N$ (*Caesar's city-destruction*).

Table 1. ***Five types of internal structures and the hapax rate of each type***

Type		Structure	Example	Rate	
I	a.	(Det) + derived N	*(the) destruction*	28.6%	32.2%
	b.	(Det) + derived N + N	*(the) destruction process*	3.6%	
II	a.	(Det) + derived N + of + NP	*(the) destruction of the city*	12.9%	16.5%
	b.	possessive + derived N	*(the) city's destruction*	3.6%	
III	a.	(Det) + $[N + V_{suf}]_N$	*(the) city-destruction*	35.9%	49.6%
	b.	(Det) + $[N + V_{suf}]_N$ + N	*(the) city-destruction process*	13.7%	
IV		possessive + derived N + of + NP	*Caesar's destruction of the city*	1.2%	
V		possessive + $[N + V_{suf}]_N$	*Caesar's city-destruction*	0.4%	

Type I is a derived noun optionally accompanied by a determiner like *the*, *this*, and *any*. This type includes the pattern "(Det) + derived N + N" (*(the) destruction process*), where a head noun is modified by a derived noun. Type II nominals include an argument; in most cases the argument corresponds to the direct object of the base verb, but it sometimes corresponds to the subject, as in *reinflation of the balloon*.[5] Additionally, what is called "passive nominal" (*(the) city's destruction*) is included in Type II. Type III consists of an optional determiner and a synthetic compound (SC), in which the second unit is deverbal and the first unit is related syntactically to the internal verb. The pattern of SC plus noun like *(the) city-destruction process* belongs to Type III. Type IV has a possessive combined with Type II composition, the possessive being interpreted as subject, and Type V is an SC version of Type IV.

The hapax rate of each type is shown in the last column of Table 1. For example, the number of Type I hapaxes comprises 32.2% of the whole hapaxes recorded. Type III is the most productive, followed by Type I. The two types make up 80% of the whole hapaxes recorded, and accordingly they are the major patterns for online formation of nominals. Note a common feature of Type I and III nominals: they are composed of a single word, a derived word (*destruction*) or compound (*destruction process*), optionally accompanied by a determiner. By contrast, Type II with one syntactically realized argument is not so fruitful, and Type IV and V hapax nominals with two arguments are attested only in very isolated coinages and can be therefore considered unproductive.

Now that the five basic types of the internal structures of nominals have been described, let us briefly examine their external behavior. The external syntax of nominal expressions can be classified into five groups, as indicated in Table 2: their use as subjects (A); their use as complements of prepositions (B), verbs (C), and conjunctions (D); and the others (E) (the ones which do not figure in larger constructions like headwords). For each group, the rate of the number of the hapax nominals identified in the BNC survey is shown in the last column of Table 2. It is clear from this result that hapax nominals can dominantly serve as complements of prepositions or verbs when they are used in the formation of larger structures.

Table 2. ***Five types of external structures and the hapax rate of each type***

Type	Syntactic Function	Example	Rate
A	subject	(1)	13.3%
B	complement of preposition	(3a)	54.4%
C	complement of verb	(ib) in note 8	26.6%
D	complement of conjunction	(9)	1.6%
E	others[6]		4.1%

3.2. Functional Properties

Let us turn next to identification of the communicative functions that are performed by hapax nominals. Three functions are particularly significant: cohesion, focus, and brevity. The first function of the hapax nominals is a role of achieving discourse cohesion. A basic internal link of discourse is thematic link in which what is first introduced as "rheme" becomes the "theme" in the subsequent text, with the theme typically associated with a pro-form (cf. Quirk et al. (1985: 1430-1431)). A good example of the thematic link by nominalization is given in (1), where the rhematic predicate "gives them a little more shut-eye" switches to a new topic in what follows and it is elegantly designated by the thematic substitute *eye-closure*. A nominal acts explicitly as a thematic substitute to contribute to text cohesion. Type I and III internal structures are optimal for substitutes and Type A external structure is optimal for a thematic expression. Thus, these structures are closely related to this function.

(1) Rather than maintain high levels of vigilance all the time, birds choose to lower their levels of vigilance when risk is low. This gives them a little more shut-eye. Why is *eye-closure* so important? (BNC B74:676)

The hapax nominals which take arguments other than direct argument are rarely attested in the corpus, and consequently attention should be drawn to the differences in information structure between nominals with and without direct object. To illustrate the differences, let us consider the three nominal patterns found in (2).

(2) a. Disclosing things about yourself often encourages other people to open up too. But in inappropriate situations and when talking with someone who is quite guarded, *over-disclosure* may be interpreted as intrusive. (BNC CEF:1400)

b. The one exception is that the opening of the inter-German frontier has complicated the campaign for the frontier-free *movement of people* … Yet, even without this complication, *people-movement* was already one of the most difficult parts of 1992. (BNC ABF:1514)

In (2a), the gerundive clause *Disclosing things about yourself* is subsequently replaced with the single complex word *over-disclosure*, with its direct object recovered from the preceding discourse. Type I nominalization therefore serves to emphasize the activity which the noun denotes and de-emphasize the object of the designated activity. In contrast, when the object of nominal emerges into the foreground of attention, Type II nominal is chosen following the principle of end-focus. The nominal *movement of people* in (2b) illustrates such a case. Further, the Type III nominal *people-movement* in (2b) represents a momentarily nameworthy category on the basis of the prior utterance. This sort of compound therefore carries out a primary function in assigning the object backgrounded information by incorporating it within the word (Rice and Prideaux (1991)).

Type I and III can be a means of representing “brevity”; by choosing a single derived noun or a synthetic compound, with its arguments syntactically implicit, we can construct concise and sensible nominals. In other words, a special conception can be produced by compressing a propositional content into a single word. Consider in this regard the following two passages. In (3a), the activity which is explicitly anticipated by the preceding clause is expressed concisely by the transiently constituted form *co-presentation*. Thus a “brevity effect” is consequently obtained from this nominalization, which is not gained from the annotating paraphrase. Passage (3b) also presents a typical example of the brevity effect.

(3) a. The tradition of the male-female presentation has been maintained. On the subject of *co-presentation*, I must come now to Richard Wyatt. (BNC EVN:1213)

b. Without the pinpoint contact of the tips of the claws, the animals may find themselves slipping and crashing to the ground. The expression of confusion observed on the faces of such cats as they pick themselves up is in itself sufficient to turn any cat-lover against the idea of *claw-removal.* (BNC BMG:673)

In sum, the main point which has to be made clear is that a major factor triggering the choice among nominal forms is closely associated with the communicative functions; an optimal nominal form is constructed in order to package the information in the most appropriate way.

4. Contextuals

The last section has discussed the internal and external structures of hapax nominals as a whole, their functions, and the relationship of the structures and functions. As suggested by Clark and Clark (1979), hapaxes, mostly a result of spontaneous composition, are generally created depending on context. In this section, we will deal with context-sensitive hapax nominals, showing that they play a core part of hapax nominalization in performing the previously discussed communicative functions, in particular discourse cohesion.

4.1. Classification of Contextuals

Three subtypes of contextuals are recognizable: (a) anaphoric contextuals, (b) cataphoric contextuals, and (c) contextuals requiring the speaker's and hearer's mutual knowledge.[7] The first type of nominal is the one which has direct connection with the item mentioned in the immediately preceding discourse. Anaphoric contextuals can be either the ones which are directly derived from the preceding VP structures as illustrated in (4), or the ones which are constructed depending on the prior correspondent phrases as illustrated in (5), where a process nominal (*centralization*) is followed by a related hapax nominal (*pluralization*).

(4) To win approval, drugs must first go through a series of animal and clinical tests which are reviewed by government *drug-approval* agencies. (BNC ABH:2094)

(5) All in all, an ever-growing institutional centralization was matched by a burgeoning theological *pluralization*. (BNC CRK:47)

The name for a certain process is normally introduced into discourse anaphorically; the process name is given after its explanation. It is possible, though, that the label is presented before its explication. The former mode may yield anaphoric contextuals as in (3a) above, while the latter way may bring forth cataphoric contextuals as in (6).

(6) The most popular *goal-arousal* theories are called "need theories." These theories collect goals, aspirations and behaviour into motives … (BNC EAA:130)

Finally, an addresser can express with a nominal the event which he/she assumes is known to the addressee. Example (7) is a case in point, where the use of the possessive *her* signals the addressee's sure knowledge of the event on the time of its utterance.

(7) But, although Margaret was now "dead," she was not yet buried. Stricken with umbrage, she had spent the months since *her destoolment* sniping at her successor … (BNC HNK:1036).

More than half of the attested hapax nominals are contextuals (135 types (54.4%)). Given that hapax nominals are online new coinages, this demonstrates that nonce nominal formation is principally made context-dependently. The number of each type is as follows: Type a (110 (81.5%)), Type b (17 (12.6%)), and Type c (8 (5.9%)). Type a is by far the most productive, which clearly shows the anaphoricity of nonce nominal formation: it plays a crucial role of affording a concise form of the propositional unit introduced into the prior text.[8]

4.2. Relationship of Contextuals to Internal and External Structures

We are now in the position of considering the context-dependency of the three main internal structures of nominals: (I) (Det) + derived N, (II) (Det) + derived N + of + NP, and (III) (Det) + $[N + V_{suf}]_N$, assuming the context-dependency (D) = N_C/N_H, where N_C is the number of contextuals and N_H is the number of hapaxes. The context-dependency of each type in the BNC survey is demonstrated in the last column of Table 3.

Table 3. ***Three main types of internal structures and their context-dependency***

Type		Structure	Example	Context-dependency (D)	
I	a.	(Det) + derived N	*(the) destruction*	0.662	0.663
	b.	(Det) + derived N + N	*(the) destruction process*	0.667	
II	a.	(Det) + derived N + of + NP	*(the) destruction of the city*	0.500	0.561
	b.	possessive + derived N	*(the) city's destruction*	0.778	
III	a.	(Det) + $[N + V_{suf}]_N$	*(the) city-destruction*	0.472	0.472
	b.	(Det) + $[N + V_{suf}]_N$ + N	*(the) city-destruction process*	0.471	

Context-dependency (D)=N_C/N_H, where N_C is the number of contextuals and N_H is the number of hapaxes.

It will be readily noticed that Type I and Type IIb (passive nominals) are highly context-dependent. This reflects the function of discourse cohesion which they carry out. Since the complement of the Type I nominal has to be retrieved for its interpretation, it naturally relies for information on the context. Further, all the passive nominals attested turned out to contain possessive pronouns (e.g. *its reidentification*), which shows that the pronouns naturally link to their antecedents for the identification of the nominal complements.

With regard to the relation of contextuals to external nominal structures, the dependency ratio in the BNC survey is shown in Table 4.

Table 4. ***Four main types of external structures and their context-dependency***

Type	Syntactic Function	Example	Context-dependency (D)
A	subject	(1)	0.636
B	complement of preposition	(3a)	0.489
C	complement of verb	(ib) in note 8	0.530
D	complement of conjunction	(9)	1.000

It should be noted that Type A and Type D are highly context-dependent, and this reflects, again, the function of discourse cohesion which they carry out. As regards Type A, there is a high dependency ratio between nominal contextuals and their use as subjects. The reason is that subjects tend to look for anaphoric

linkage to the prior syntactic materials, since subjects are unmarked topics (Lambrecht (1994: 132)) and topic expressions designate the topic referents anaphorically (Lambrecht (1994: 187)). In this connection, it is relevant to note that there is a case where a nominal with *of* serves as complement of head noun but the whole NP serves as subject, as with *this issue of root-assignment* in example (8). Eight examples of this case have been attested and seven of them are contextuals. If these are classified as external subjects, the subject use of hapax nominals becomes more context-dependent (0.683). As for Type D, the conjunctions involved are all those which introduce temporal adverbial phrases, as in (9). A temporal adverbial clause or phrase normally expresses what is known to the addressee, hence providing the temporal background for the main clause (Lambrecht (1994: 125)). Thus, anaphoric linkage is established between a temporal adverbial and the prior syntactic materials.

(8) So if "pay" and "payment" are indexed as having separate roots, the above fragment would show no overlap. However, if "payment" were assigned the same index as "pay," then a strong overlap would result. … It is suggested that this issue of *root-assignment* (or lemmatisation) and grain-size form the basis of further investigation. (BNC EES:1258)

(9) … which was situated near the church where she eventually had her cell: … But there is no evidence that Julian had ever been a nun *before her immurement* … (BNC CD4:467)

From the above observation, we are justified in asserting that contextuals considerably reflect how closely the structural properties of nominals are bound up with a function of nominals—discourse cohesion.

5. Conclusion

We have conducted an in-depth analysis of the hapax nominals identified in the BNC survey. Some creative aspects of online nominalization have been disclosed: its optimal inner- and outer-structures, context-dependent communicative functions, and the close interrelationship of the forms and functions. It is worth further investigation to see how and to what extent unstocked creative kind of nominalization is related to stocked uninventive kind.

Notes

* This article is a revised version of the paper read at CBA 2010: Workshop on Corpus-Based Approaches to Paraphrasing and Nominalization held at Universitat de Barcelona on December 1, 2010. This work is partly supported by a Grant-in-Aid for Scientific Research (C) (No. 26370462) from the Japan Society for the Promotion of Science.

[1] As seen in example (i), the derived nouns *relativisation/relativization* and *problematisation/problematization* are considered as neologisms or words not previously encountered by a writer although the words have a frequency of 4 occurrences and a frequency of 13 occurrences in the BNC survey, respectively. Since even a neologism can be used more frequently than once in a large-scale corpus, a hapax nominal, which is of token frequency 1, is highly qualified for "neologismhood."

(i) Neologisms such as "pluralism", "simulationist", "*relativisation*", "*problematisation*" imply a world in which certainty has been cast into abeyance. (BNC FBF:35)

[2] For hapax-finding I am indebted to the research engine of http://view.byu.edu/ reg3.asp?c =aybfiyfml.

[3] Chapin (1967: 13-17) proposes that "self-ing" derivatives are derived by a transformational rule.

There is one drawback with our research engine: hyphenated compounds (*city-destruction*) can be found by it, but non-hyphenated compounds (*city destruction*) cannot. To reduce the drawback, we have checked whether there is a non-hyphenated counterpart for a hyphenated hapax compound. If such a counterpart is detected, the hyphenated hapax does not count as true hapax. For instance, the BNC contains not only the compound *land-allocation*, which is of token frequency 1, but also the non-hyphenated counterpart *land allocation*, and hence *land-allocation* is judged to be a non-hapax here.

[4] It cannot be denied, however, that some derived nouns which resist reorganization might be contained in these hapaxes; derivatives like this may tend to have monomorphemic base verbs which are not compositionally interpretable (e.g. *dimidiation* and *crepitation*).

[5] The *Theme* direct argument of a nominal head is generally marked by *of*, but in some cases it is marked by some other preposition as in *over-insistence on bureaucratic process*. The latter case as well as the former one is included under Type IIa.

[6] Type E includes a kind of headword, as in:

(i) The continued expansion of the S.M.E. Centre's activities has created excellent opportunities for high caliber staff in the following posts: Lecturer in Small and Medium-Sized Enterprises Management (*Re-advertisement*) (BNC CJU:384)

[7] The classification of Type c is due to Clark and Clark (1979).

[8] Since context-dependency is not a required condition for online nominalization, there is a considerable number of non-contextual hapax nominals (113 types (45.6%)). Two points are worth noting here. First, among the non-contextuals are those whose corresponding verb is widely used or comparable verb-complement combination is well established. For example, for the hapax nominal in (ia) the underlying verb is commonly used, and for the hapax nominal in (ib) the underlying verb-complement combination is relatively well attested and semantically transparent, so that the hapaxes at issue are easily interpretable without contextual force.

(i) a. … that would allow you to tell complicated stories simply for the aesthetic pleasure of complexity of complication and *unravelment*, suspense, and the rest. (BNC APS:527)

b. They are planning another *land-occupation* for next weekend, about two kilometers from here. (BNC CAH:602)

Second, a significant proportion of non-contextuals occur in coordinate constructions (30 types (26.6%)), as in (ii) (See also (ia)). The syntactico-semantic similarity of coordinates seems to provide an aid to the interpretation of a non-contextual occurring as coordinate.

(ii) ATV had run out of cash, as a result of large start-up costs, a delayed opening of its Midlands service and general *under-capitalization*. (BNC CRY:1190)

References

Allen, Margaret R. (1978) *Morphological Investigations*, Doctoral dissertation, University of Connecticut.

Aronoff, Mark (1980) "Contextuals," *Language* 56, 744-758.

Baayen, R. Harald and Antoinette Renouf (1996) "Chronicling *The Times*: Productive Lexical Innovations in an English Newspaper," *Language* 72, 69-96.

Chapin, Paul G. (1967) *On the Syntax of Word-Derivation in English*, MITRE, Bedford MA.

Clark, Eve and Herbert Clark (1979) "When Nouns Surface as Verbs," *Language* 55, 767-811.

Hay, Jennifer (2003) *Causes and Consequences of Word Structure*, Routledge, New York.

Jackendoff, Ray (1997) *The Architecture of the Language Faculty*, MIT Press, Cambridge, MA.

Lambrecht, Knud (1994) *Information Structure and Sentence Form*, Cambridge University Press, Cambridge.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik (1985) *A Comprehensive Grammar of the English Language*, Longman, London.

Rice, Sally and Gary Prideaux (1991) "Event-Packing: The Case of Object Incorporation in English," *BLS* 17, 283-298.

Corpus and Dictionaries

The British National Corpus.

The Oxford English Dictionary on CD-ROM (Version 2.0), 1999, Oxford University Press.

Reverse Dictionary of Present-Day English, ed. by Martin Lehnert, 1971, VEB Verlag Enzyklopädie.