Keita Tsuji, Graduate School of Education, University of Tokyo, Japan
Kyo Kageura, University of Sheffield and NACSIS, UK and Japan

ANALYSIS OF WORD STRUCTURE OF MEDICAL SYNONYMS

1. Introduction

In accordance with the growth of technical terminology, the number of synonyms also grows. When the concept referred to be synonyms become stable, usually one term becomes dominant among the synonyms. Then what kind of term becomes dominant? Few work has been done in this respect. It is said that language changes mostly by chance, but there may be some regularities in the choice of the dominant term among scientific synonyms. Such regularities, if we can find them, are not only interesting as a theoretical study of synonymous terms but also practically useful in selecting headings or descriptors of technical dictionaries or thesauri.

Against this background, we examined a number of synonymous terms, from four points of view, i.e. the number of mora (phonetic term length of Japanese, origins of terms or 'gosyu' (Chinese origined, Western-language origined, original Japanese, and their mixture), semantic structure of terms as represented by constituent units, and user group of synonyms. Intuitively, the number of mora is related to the physical cost necessary for using the term, the origin is related to the stylistic and communication efficiency, the semantic structure is related to the understandability and systematicity of terms among the overall terminology of the field, and user group is a measure for authority given to the word.

Here we would like to report the result of the analysis of the synonymous terms from the point of view of their semantic structures. The terms analyzed are terms which represent various diseases. As far as Japanese is concerned, there are many synonymous terms representing diseases. Of course the tendency observed in the choice of dominant term among synonyms may differ from one field to another, and from a concept types to another. So we do not claim that the result reported here is valid in general. We believe a number of this sort of concrete study is necessary before we can draw any general conclusions about the regularities of process of dominant term choice among synonymous terms.

2. Data Preparation

2.1 Extracting Synonyms

Though the definitions of synonym differ from one scholar to another and some even claim that there is no word pairs which are synonymous. However, there are some pairs which can intuitively regarded as representing same concepts. In choosing synonymous pairs for our analysis, we take a practical standpoint and regard the terms indicated as synonyms in standard terminological reference tool as synonyms.

TKE'96, Terminology and Knowledge Engineering. Frankfurt: INDEKS-Verlag 1996. p. 190-196

The reference tool we used is NANZANDO'S MEDICAL DICTIONARY (NMD). It clearly shows synonyms as well as the preferences among the synonyms. For instance, in NMD, the entry 'Basedow disease' is followed by '((Graves disease))' and at the entry of 'Graves disease', 'Basedow disease' is referred to by arrow, i.e.'--> Basedow disease'. We regard that the term which is referred to by other entry as the dominant term among synonyms. Thus in the above example, 'Basedow disease' is the dominant synonym over 'Graves disease'. By the way, for the convenience's sake we would henceforth call the term whose entry has bracketed synonym under it as 'main term', and the bracketed synonym which refer to the main term as 'synonymous term'. The number of main term is 428, and the number of synonymous term is 600. The latter is bigger than the former because some main terms have more than one synonymous terms.

2.2 Extracting Constituent Units of Terms

For analyzing semantic patterns of terms, it is necessary to decompose the terms and extract basic constituent units. We decomposed the terms into constituent unit roughly following the guideline of extracting "Beta Unit" established by National Language Research Institute. Beta unit is the most frequently used unit so far in large scale studies of lexicology, and therefore considered to be standard for Japanese. Roughly speaking, the constituent elements extracted according to Beta unit criteria consist of one and two Chinese character units, Katakana units which is roughly equivalent to English simple words, and simple Kana words.

3. Recognition of Semantic Categories

We examined the constituent units of terms representing diseases and recognized two types of semantic categories. The first is based on the concepts represented by individual constituent units, while the second is based on the function or role of the constituent units within the terms. So the former is conceptual categories, while the latter is functional categories or relational categories. This distinction corresponds to the distinction of intra-term relationships and conceptual categories of constituent units, though we decided to analyze intraterm relationships not as the relationships between constituent elements but as the relationship of each constituent element to the term as a whole. This is why we attribute functional categories to constituent elements. The conceptual categories we established are as follows.

	Conceptual Category	Example
1	proper noun of person	Basedou(Basedow)
2	proper noun of place	Berurin(Berlin)
3	virus or germ or insect	barutonera(bartonella)
4	creature except people,germ,insect	kame(turtle)
5	expression connected with heredity	senten(congenital)
6	expression connected with sex	joshi(female)
7	class of age	yoji(infant)
8	occupation	sensuihu(diver)
9	physical component of human body	i(stomach)

10	mind or sense	shikaku(sense of sight)
11	quantity of human body	shincho(height)
12	point of time	jutsugo(after operation)
13	period of time	mittsuka(three days)
14	stage	shoki(early)
15	curability	ryosei(benign)
16	speed	kyusei(acute)
17	frequency	kanketsusei(intermittent)
18	formal classification	A gata(type A)
19	position or place	migi(right)
20	color	kokushoku(black)
21	shape	kyujo(globular)
22	volume	kyodai(giant)
23	area	buroodo(broad)
24	density	sosho(low density)
25	weight	juryo(heavy)
26	temperature	kanrei(cold)
27	humidity	kanso(dried)
28	solidity	nansei(soft)
29	force	koatsu(high pressure)
30	number	niju(double)
31	action	shindo(vibrate)
32	expression connected with disease	byo(disease)
33	object or substance which is not included in above categories	hikoki(air plane)

Table 1. Conceptual Categories

On the other hand, we recognized seven functional categories, i.e.:

Finder of disease (F), Place where disease prevailed (L), Cause of disease (C), Patient of disease (P), Part of human body or function at which the disorder occurs (H), Simili expression which characterizes the disease (M), Non-simili expression which characterizes the disease and is not included in above categories (D).

Some of these functional categories have strong correspondence to the 33 conceptual categories. For instance, "Cause of disease" corresponds to 7 conceptual categories, i.e. 3, 4, 5, 26, 27, 29, and 31. There are, of course, some one-to-many correspondences between these two types of categories. For instance, a unit which belongs to the conceptual category 33 plays different roles in different terms.

4. Results

Using the above types of categories, we classified the sample and counted the pairs which include these units. The result is shown in table 2. The "only main" column shows the number of pairs where only the main term includes the unit indicated in the first column, while the "only syn" column shows the number of pairs where only the synonymous terms include the category. The column of "main & syn" shows the number of pairs whose main and synonymous terms both include the category.

Comparing the number of "only main" with that of "only syn", we can observe what types of terms are dominant from the point of view of conceptual and functional categories which the terms include. The categories whose numbers of "only main" and "only syn" are significantly different (at 0.05 percent level) are: F1, L2, C3, C4, C31, C33, P7, H11, D15, D24, D31. They are indicated by '*' in table 2.

	Functional Category	Conceptual Category	only main	only syn	main& syn	difference
F	Finder of the disease	1	65	91	23	*
L	Place where the disease prevailed	2	1	7	0	*
		33	3	8	5	
C	Cause of the disease	3	15	5	8	*
		4	8	1	3	*
		5	25	16	20	
		9	0	0	0	
		26	0	1	2	
		27	0	1	0	
		29	0	1	0	
		31	3	0	6	*
		33	12	2	3	*
P	Patient of the disease	6	0	0	0	
		7	5	21	10	*
		8	2	3	1	
H	Part of human body or function at which the disease occurs	9	81	90	246	
		10	2	0	0	
		11	0	3	0	*
		31	16	13	29	
M	Simili expression which characterizes the disease	4	11	6	4	
		19	0	0	0	
		31	2	2	5	
		32	1	0	0	
		33	5	3	4	
D	Non-simili expression which characterizes the disease and is not included in above categories	12	1	2	5	
		13	4	2	0	
		14	1	3	1	
		15	1	6	1	*
		16	4	2	3	
		17	0	2	1	
		18	10	10	0	
		19	0	1	0	
		20	13	11	14	
		21	7	5	4	
		22	3	3	9	

		23	0	0	0	
		24	3	0	0	*
		25	0	0	0	
		26	0	0	0	
		27	0	0	0	
		28	1	0	2	
		29	0	1	0	
		30	0	1	1	
		31	81	113	76	*
		32	22	24	549	
		33	7	14	27	

Table 2. Categories and Main/Synonymous Terms

At the level of functional categories, on the other hand, it is observed that the categories F, L, C, and P shows significant difference between the numbers of "only main" and of "only syn" (at 0.05 percent level). The term containing C is likely to become the main term. Those containing F, L, and P are likely to become synonymous terms. These are based on conceptual categories. We also examined each pair of the functional categories to see mutual dependency of the functional categories.

For example, 'Kuron byo (Crohn's disease)' and 'shumatsu kaicho en (terminal ileitis)' are synonymous. And the former contains category F and the latter contains H, reversely the former does not contain H, and the latter does not contain F. 'Hipperu byo (von Hippel's disease)' and 'momaku kekkan shu sho (retinal angiomatosis)' are the same. In these cases, if the latter were likely to be the main term, it could be said that the term containing H is likely to be dominant over the term containing F. In this view, we examined the sample and got the following result. X-axis is the main term and Y-axis is the synonymous term.

	F	L	C	P	H	M	D
F	-	0	25	5	51	3	0
L	0	-	10	0	3	0	0
C	11	1	-	1	2	1	0
P	10	2	9	-	1	0	2
H	50	1	11	2	-	4	1
M	2	0	1	0	4	-	1
D	1	0	3	0	0	0	-

Table 3. Combination of Categories and Main/Synonymous Term

In this table 3 we can see that the number of pairs <<one of which contains F and does not contain H>> and <<the other of which contains H and does not contain F>> is 50+51 = 101. The number of pairs at which the former is the main term is 50, and the number of pairs at which the latter is the main term is 51. There is no significant difference between them and we can't say which kind of term is more likely to be the main term containing F or H. But there are significant difference (at 0.05 percent level) within the pairs (F,C), (L,C), (P,C), (H,C), (D,C). In all of them, the term containing C is more likely to be the main term, and in that sense, likely to be dominant.

5. Discussion

In table 2, we saw that the term which does not contain any of F, L, P is likely to become dominant over those which contain these categories. But table 3 tells us that those results are attributed to their combination with C because significant difference are not seen in case of other combinations. So it is worthwhile to consider why terms containing C are likely to become dominant over other terms.

Firstly, many diseases are named after the person who made the existence of the disease known, i.e. the term with the category F. These terms typically take the form of "(name of person)'s disease". These terms, however, are not very informative, because no inherent characterizations of the disease are expressed in the terms. In addition, it sometimes happens that the same disease is named after different persons as it is 'found' independently by different persons. We can see many such synonymous pairs, e.g. "Milroy's disease" and "Meige's disease". In case of the terms with place name (L), the situation is quite similar to those with F. The place where the disease first or most notably prevailed is not so informative to be included as part of the term. In addition, the same disease may prevail in different places. In Japan, the same disease have three different names, all of which are named after the location where the disease prevailed, i.e. 'Yamanashi byo (Yamanashi disease)', 'Katayama byo (Katayama disease)', 'Saga ryuko byo(Saga prevalent disease)'. It is only later that they are found to be the same disease, caused by the same Trematoda, Schistosoma japonicum.

In short, the names after founder or place not only lack inherent information of the diseases but also can be misleading. Compared to these, the cause is considered to be an essential characteristic of disease and thus gives a useful information if it is included in the term. Finding the cause also have some impact on medical scientist, or the system of concepts. Because it sometimes makes clear that some diseases which are considered to be different is in fact the same disease. That is a unification of concepts, and in that case, the term containing cause is accepted and favored as a kind of unifier among medical scientists. Its neutral or a kind of superordinate impression seems to help it be dominant synonym among person's or place's name disease synonyms. Many terms contain the units which represent parts of body, many of which also contain the units representing trouble or disorder observed on the parts. Like the cause of disease, they are also considered to represent useful information. However, some diseases accompany many disorders, none of which is dominant. For example, bartonellosis which is caused by microbe bartonella, accompanies high fever, headache, anemia, arthralgia, warts, swelling of liver/spleen/lymph node. So if we name a disease using only expressions of disorder and the disease has no characteristic disorder, we must combine many to make it distinguishable. On the other hand, different diseases may have the same characteristic disorder, which is also confusing.

6. Concluding Remarks

We saw that the term which contains the cause of disease is likely to be dominant over other disease names. This phenomenon can be rationalized from the point of view of the naming principle that the useful and essential information which characterizes the concept to be represented must be used for naming. Although it is only a specific instance of the naming principle which is clarified in the present examination, we believe this sort of individual analysis is needed before we establish the concrete naming principle of a field, not to speak of the naming principle in general.

References

CRUSE, D.A. (1986): Lexical Semantics. New York: Cambridge University Press.

SPARCK-JONES, K. (1986): Synonymy and Semantic Classification. Edinburgh: Edinburgh University Press.

LYONS, J. (1981): Language and Linguistics - An Introduction. Cambridge: Cambridge University Press.

NIDA, E.A. (1975): Exploring Semantic Structures. München: Wilhelm Fink Verlag.

IKEGAMI, Y. (1975): Semantics: Analysis of Semantic Structures (in Japanese). Tokyo: Taishukan-shoten.

NANZANDO (1990): Nanzando's Medical Dictionary (in Japanese). Tokyo: Nanzando.