Vesa Kautto
Department of Library and Information Science
University of Oulu, Oulu, Finland

Classing and Indexing: A Comparative Time Study

Kautto V.: **Classing and indexing: a comparative time study.** Int.Classif. 19(1992)No.4, p.205-209, 17 refs.
A total of 16 classifiers made a subject analysis of a set of books such that some of the books were first classified by the UDC and then indexed with terms from the General Finnish Subject Headings while another set were processed in the opposite order. Finally books on the same subject were either classified or indexed. The total number of books processed was 581. A comparison was made of the time required for processing in different situations and of the number of classes or subject headings used. The time figures were compared with corresponding data from the British Library (1972) and the Library of Congress (1990 and 1991). The author finds that the contents! analysis requires one third, classification one third and indexing one third of the time, if the document is both classified and indexed. There was a plausible correlation (0.51) between the length of experience in classification and the decrease in the time required for classing. The average number of UDC numbers was 4.3 and the average number of terms from the list of subject headings was 4.0. (Author)

1. Introduction

Classification systems and indexing languages have been compared in many ways. Their structural properties have been analysed, and it has been found that faceted classification and a thesaurus are closer to each other than enumerative classifications and subject headings (1). This closeness is also indicated by the instruction given nowadays for thesaurus construction (2). Classification and indexing systems or documentation languages have been examined from a linguistic point of view by Hutchins (3), for example, who states that it is possible to speak of their vocabulary, grammar and semantics. Hutchins has also drawn attention to the paradigmatic and syntagmatic features of documentation languages. The philosophical-linguistic examination of classifications and indexing languages by Svenonius is interesting, although I do not agree with all the conclusions she draws (4).

Comparisons have been made between UDC and subject heading lists/thesauri a.o. with regard to the number of descriptors and their adequacy. The retrievability of references when UDC and other search languages are used has been measured experimentally in the Aslib-Cranfield project, for example, and in tests arranged by CIINTE and VINITI during 1957-1970. The transformation of UDC into a subject heading list has also been studied (5), while Dahlberg (6) has analyzed UDC from the point of view of an ideal documentation language.

Literature comparing the time required in classing and indexing is scarce. In this respect Reynolds' comment in 1973 still holds good. She went through 130 texts on the topic, 28 of which contained information about the cost and duration of subject analysis: She states: "As a result of the literature search, we are confirmed in our previous opinion that cost data in this area is scarce, difficult to evaluate and contradictory. The literature abounds in opinion, but hard data, together with detailed description of the content in which it has been produced, is extremely elusive" (7). A further problem is expressed by Line: "Most studies of catalogue costs have taken into account both cataloguing and classification, and have not separated the two" (8).

The largest studies on the time required for classing and indexing are those carried out in national libraries. In 1972 the British Library measured the time required for classing (DDC, LCC) and indexing (PRECIS, BMSI and LCSH); the analysis covered tens of thousands of titles (9). I have similar data for the Library of Congress, based on the whole year's work, of the time spent in classing and indexing (LCC, LCSH and DDC) during the years 1990 and 1991 (10). These studies yielded the following results.

Table 1. Time required for classing and indexing in the British Library in 1972 (9):

Class. scheme	Units processed	Units/indexer/day	Time/unit in minutes
DDC	17800	37	10
LCC	17800	49	7
Index. scheme			
BMSI	58000	54	7
LCSH	17800	49	7
PRECIS	33000	34	11

It is worth noting that in the British Library indexing with PRECIS preceded indexing with LCSH and classing with LCC and DDC.

Table 2. Time required for classing and indexing in the Library of Congress in 1990 and 1991 (10):

	Titles/hour		Time/title in minutes	
	1990	1991	1990	1991
LCC and LCSH	2.23	2.22	27	27
DDC	7.84	7.94	8	8

Data on the time required for classing with UDC is scarce. A minor study (148 titles) is mentioned by Line, in which the average time was 4.6 minutes per title (8). Nurminen measured the time spent in classing by UDC in the Joensuu University Library. In 1976 the number of titles processed was 6000, and the average time spent in classing was 11 minutes per title (11).

There seems to be virtually no comparison of subject analysis using both UDC and indexing on the same material. This kind of study I shall attempt to explain in the following. My investigation had a practical goal. In Finland, UDC has been the most common means of subject analysis used in research libraries for several decades (12). Its use in subject retrieval, however, has proved to be slight (13). In 1988 the first experimental version of the General Finnish Subject Headings (GFSH) covering all fields of knowledge was issued (14). It contains about 11000 subject headings which can be linked to each other by means of colons to form a chain. In addition the list contains more than 1900 see references. At the University of Oulu Library consideration has been given to the adoption of this list of subject headings alongside UDC or instead of UDC for the subject analysis of books. In both cases it is important to know, whether the subject headings are able to describe the literature acquired in a large multidisciplinary library, which classifies about 11000 foreign and 5000 Finnish titles annually. If the subject headings are adopted alongside the UDC, it is important to know, how much additional work this would entail. In the situation where UDC is to be replaced with another system, the amount and ratio of the required work load are also of interest. Corresponding questions are also being deliberated in other libraries in Finland. My investigation might, therefore, be of significance in quite far reaching decision making.

The experiment does not lack theoretical interest, either. Beghtol (15) referring to the text linguist van Dijk is of the opinion that a two-way interaction of the same sort as when reading and processing text occurs, when the contents of a document are analyzed. The reader is continuously making hypotheses with regard to the topic and meaning of the text and adopting his knowledge structures to the new information. The person performing the subject analysis must also consider the terminology and structure of the documentation language he uses and examine the document through it.

If the same person carries out the subject analysis of a document using two documentation languages consecutively, it can be assumed that a specific amount of time goes first in defining what the document is about and what concepts should be included in the analysis. The remainder of the time is then spent on the selection of corresponding class numbers on the one hand and subject headings on the other. In this test situation these periods of time can be measured.

A theoretical approach can also be applied to the number of descriptors used. If the descriptors assigned first are merely translated into the other language, it can be assumed that class numbers and subject headings describing one document will be about equal in number. If, however, the describing language has a direct influence, it can be assumed that the number of descriptors will vary.

2. The Experiment

The experiment was carried out around the turn of the year 1989/90. 16 librarians and information specialists from 10 different units of the University of Oulu Library took part. Of these three classifiers from the library of the Faculty of Education worked and reported as a team, and therefore the number of testers can be cut down to 14. The majority of them had university degrees in the field they dealt with. For instance, the classifier in the Faculty of Technology was a graduate engineer, and that in the Biology Library a botanist. Experience in UDC classification varied very much. Five classifiers had less than one years experience, four 1-5 years, and seven 6-20 years. The Finnish abridged edition of the UDC was used for the classing, supplemented with the complete English version for the subject in question.

The librarians and information specialists participating in the experiment were, naturally enough, less experienced in the use of the GFSH. Only three had used it previously, while six others had experience in the use of other subject heading lists varying from four months to seven years.

It can be considered a shortcoming of the experiment that the participants were not given uniform training in indexing, although they became acquainted with the subject heading list in the course of the planning of the experiment. The use of the subject heading list was learnt by means of its clear directions for use. A thorough familiarization with GFSH would have required more time, than could have been organized within the framework of the experiment. According to Lufkin, full command of indexing takes three months (16).

The test instructions requested that each participant should process 20 (or 10) books first classifying them and then indexing them (group A), then again 20 (or 10) books first indexing them and then classifying them (group B). In addition I asked each person to select 10 books for classing only (group C) and 10 books which, judging from the title, seemed to be similar to the first, for indexing only (group D). For each book the following information was recorded on a form:

- title of book
- UDC numbers
- subject headings
- relevant subject headings not included in the list of subject headings
- the time required for classing and indexing
- an appraisal of how well the title described the contents.

The participants also prepared general statements comparing the two systems with each other.

Altogether 581 books were processed. They were divided in the above groups as follows:

A: 178
B: 169
C: 116
D: 113.

47 % of the books were published during 1989 - 1990. 70 % of the books were in English, and less than 10 % were in Swedish, German or Finnish. The distribution by subject was fairly even, books on the sciences and arts were most common, after which came books on the social sciences, technology and medicine. Each participant processed on average 41.5 books.

The experiment was carried out in a natural situation, most participants had to interrupt processing and deal with customers etc. in between. The time recorded was, naturally, the actual time taken for classifying and indexing.

3. Results

3.1 Processing time

Of particular interest is the wide scatter of processing times as illustrated in Table 3.

Table 3. Minimum and maximum processing times in minutes

Group	Classing		Indexing	
	minimum	maximum	minimum	maximum
A	1	105	1	50
B	0,5	55	1	60
C	1	70	-	-
D	-	-	1	65

There are considerable differences in the maximum times between the individual participants. This aplies to both classing (7-105 minutes) and indexing (12-65 minutes).

Table 4 shows the average and median times required for subject analysis. Medians are given in addition to the average figures because of the wide scatter of individual times.

Table 4. Means and medians of processing times in minutes

Group	Mean		Median	
	Classing	Indexing	Classing	Indexing
A	12,0	6,4	7	5
B	7,2	11,4	5	8
C	11,3	-	6,5	-
D	-	12,0	-	8,5

On the basis of averages the time required for subject analysis when carried out as the first measure or the sole measure is 11-12 minutes. Secondary subject analysis requires 6-7 minutes or more than half of the time needed for the primary analysis. Accordingly it could be concluded, that on average, 6 minutes is needed for getting acquainted with the contents of the document, while the selection of class numbers and subject headings requires about the same amount of time. When examined on the basis of medians the picture is different. It seems that outlining the contents of the document takes 2-3 minutes, and classing or indexing about five minutes.

When the mean values are examined participant by participant, it is seen that within task A four participants spent more time on secondary indexing than on primary classing. For two of the participants this finding can be explained by a very long and thorough experience in classification and a complete lack of experience in indexing. When the whole test group is considered the first subject analysis of a document required more time at a risk level of 0.01 % than the second, if both classing and indexing was carried out.

The time spent on classing is close to the result obtained in Joensuu (average time per document 11 minutes) which renders credibility to the results obtained in Oulu. If the Library of the University of Oulu were to take up indexing in addition to classification, the working time required for subject analysis would, of course, increase. Since subject headings for the greater part (about 60 %) of the Finnish literature are readily available in the national bibliography database, this increase would be about 40 % if calculated on the basis of average times and 57 % if calculated on the basis of the median times. It can also be assumed that indexing would become less time consuming as work experience grew. Experience in classing showed a rather good correlation (0.51) with the decrease in time spent on classing, while a somewhat weaker correlation was found with the decrease in the time spent on indexing (0.30). Experience in indexing had only slightly decreased the time needed for indexing (correlation 0.16).

Classing and indexing times were also examined by shelf class. Both in classing and in indexing mathematics (average classifying time 26 minutes, indexing 23 minutes) and physics and chemistry (classing on average 17 minutes and indexing 13 minutes) needed the longest deliberation times. Later it was found that the greatest inadequacies in the GFSH when compared to the UDC had been experienced with regard to these subjects.

3.2 The Number and Adequacy of Descriptors

The number of UDC numbers assigned and the number of subject headings were counted as well as the number of missing subject headings. Each individual UDC number (main numbers, general auxiliaries and most special auxiliaries) was considered as a separate number. The same number or subject heading, when repeated, was recounted each time it appeared.

The greatest number of UDC numbers given to any one document was 18 and that of subject headings 14. The average number of UDC numbers per document was 4.3 and the average number of terms taken from the GFSH was 4.0. When those subject headings indicated are missing are included, the average number of subject headings becomes 4.5. This indicates that subject analysts are satisfied with the same, rather limited analysis both when classifying and when indexing.

The participants recorded those cases where the accuracy of the term in the subject heading list was inadequate or missing altogether. An adequate accuracy was reported for 83.7 % of the terms used. The classifiers of the Library of Physics and Mathematics expressed the greatest dissatisfaction followed by those in the Library of the Faculty of Technology. However from a more rigid point of view, it could be said that an item could not be indexed using the GFSH, if even a single additional term was needed or a single term from the list was considered inaccurate. This viewpoint gives a figure of 62.3 % for documents which could be analyzed in a satisfactory way using the subject headings. Perhaps it can be considered a good accomplishment, when a 10000 term list proves to be this adequate. Subject headings were missed most often in social sciences, mathematics, physics, chemistry and the arts.

The general evaluation was that it was easier to use the subject heading list and that it, as a recent product, was more up to date than the UDC. Indexing with subject headings was considered in some subjects (e.g. in literary research) to lead to longer and clumsier chains than classifying did. It was generally felt that more hierarchy and connections were needed in the GSFH. At present it contains only micro hierarchies in connection with terms and a very broadly grouped systematic section. In physics and mathematics the Finnish terminology in new fields of research was found to be inconsistent or nonexistant. In some other fields (biology, geology) the subject headings proved to meet the indexing needs surprisingly well.

The following evaluation was given concerning the ability of the title of the document to describe its contents. In 47 % of documents the title was found to describe the contents very well, in 31 % rather well and only in 5 % inadequately. This means that an online search of documents about a given subject using words in the title has a fairly good chance of success, if the searcher has a good command of the languages used and sufficient inventiveness. The information content of the titles was considered least adequate, by certain classifiers in the arts and by one technology classifier.

4. Conclusion

The average times required for classing and indexing that were obtained in the study were very close to those obtained by The British Library and the Joensuu University Library. The primary content analysis (PRECIS) took 11 minutes in the British Library; while exactly the same time was used the primary UDC classification in Joensuu. In my study the average time for the primary processing varied between 11.3 and 12 minutes, the median time was shorter (6.5 - 8 minutes). The average time spent on classifying with DDC, measured by the Library of Congress, is also quite close (8 minutes).

My results and those of the British Library are close if we look at the reduction of time that occurs if the primary content analysis has already been done with another subject analysis method. Secondary indexing with DDC was only one minute shorter, but with LCC and LCSH it was four minutes shorter than with PRECIS. My results are rather close to four minutes.

So it may be concluded that if a document is both indexed and classified, about one third of the time is spent in determining the subject of the document, after which choosing the classes and subject headings both take one third.

If an individual library is deliberating as to whether to use both classification and indexing in subject analysis, the following counsel could be offered. It might prove difficult to arrange for an increase of 40-60 % in the time required despite the fact that calculations seem to suggest that classing takes up a very small part of the total work load of the staff of Oulu University Library. For this reason it can be recommended that the library simplify classification if indexing is adopted as a supplement to classing, so as not to increase the total work load. It might also be worth while considering the adoption of indexing alone and the translation of the classes of literature acquired earlier into subject headings (17).

Even if classing and indexing together took the same time as in the Library of Congress (27 minutes per title), this would not mean more than 6 man years in the Oulu University Library, that is 6% of the total work time of the library. This implies that 16 000 titles per year are processed, the effective work day is 6 hours and there are 220 work days in a year. As at least the subject analysis data for domestic literature is supplied by others, the work required is even less.

It can be said that the contents of books are analyzed in the library in a rather limited manner. An average of 4 classes, of which some are auxiliaries or special auxiliaries, is not much for a document. 7-12 minutes of working time to analyze a document is not such a long time. A deeper and more diversified analysis, however, need not be aimed at, if the result of the analysis is not utilized when searching the literature. Observations of this use can be made by analyzing transaction log files of online catalogs.

More generally, it can be stated that library staff with experience of classing rapidly learn to use indexing with subject headings, and that a long experience in classing also speeds up indexing.

Acknowledgements

I would like to thank the classifiers of the Oulu University Library for their cheerful cooperation, Mr. Janne Himanka for the statistical processing of the findings, and Ms. Kyllikki Kemppainen for her assistance in the coding of forms and in other tasks. I also thank Mr. Robert M. Hiatt for providing me with time and cost figures of classing and indexing in the Library of Congress. Dr. Kalervo Järvelin, Associate Professor, has made several comments on my text, this time more beningly than usual.

References

(1) Toman, J.: Vergleich der modernen alphabetischen Ordnungssysteme und Klassifikationssysteme. In: Entwicklung von Deskriptorsystemen und ihre Nutzung beim Wiederauffinden von Informationen. RWG -Symposium. (ZIID-Schriftenreihe 12.) Berlin: Zentralinstitut für Information und Dokumentation 1966. p. 17-32.

(2) Aitchison, J., Gilchrist, A.: Thesaurus construction: A practical manual. 2nd. ed. London: Aslib 1987. 173 p.

(3) Hutchins, W. J.: Languages on indexing and classification. A linguistic study of structures and functions. Stevenage: Peter Peregrinus 1975. 148 p.

(4) Svenonius, E.: Indexical contexts. In: Universal classification I. Subject analysis and ordering systems. (Studien zur Klassifikation 11.) Frankfurt: Indeks Verlag. p. 125-138.

(5) Scibor, E.: UDC in relation to thesauri: a state-of-the-art report. In: New trends in documentation and information. Proc. 39th FID Congress. University of Edinburgh, 25-28 Sept. 1978. London: Aslib 1980. p. 248-258.

(6) Dahlberg, I.: The UDC and an ideal indexing language. In: Proceedings of the International Symposium "UDC in relation to other indexing languages". Herceg Novi, June 28 - July 1, 1971. Belgrad: Yugoslav Center for Technical and Scientific Documentation and International Federation of Documentation 1972. p. 1-25.

(7) Reynolds, R.: Literature survey on time and cost data for classification and indexing. March 1973. In: British Library Working Party on Classification and Indexing. London: British Library 1975. (BRLD report 5233). Appendix OC Class (73):15. 15 p.

(8) Line, M. B.: The cost of classification: a note. Cat. & Ind. 16(1969) Oct., p. 4.

(9) Crews, A. D.: Time and cost data for subject indexing and classification in the British National Bibliography, the Science Library and the British Museum Library. August 1973. In: British Library Working Party on Classification and Indexing. London: British Library 1975. (BRLD report 5233). Appendix OC Class (73):22. 11 p.

(10) Hiatt, R. M., Assistant to the Director for Cataloging, Library of Congress: Letter to the author Sept. 18, 1992.

(11) Nurminen T.: Kokemuksia luetteloinnista Joensuun korkeakoulun kirjastossa. (Experiences of cataloging at the Joensuu University Library). Signum 10(1977) No. 1, p. 3-7.

(12) Haarala, A.-R.: The role of UDC in Finnish classification policy. Int. Cat. & Bibliog. Control (1991) July/Sept., p. 43-46.

(13) Hakala, J., Piukkula. J.: Niin metsä vastaa - Miksi UDK-haku on harvinaisuus (The answer is an echo - Why is a UDC search a rarity?) Signum 25(1992) No. 2 p. 38-40.

(14) Yleinen suomalainen asiasanasto. (General Finnish Subject Headings). Helsinki: Helsinki Univ. Library 1988. 427 p.

(15) Beghtol, C.: Bibliographic classification theory and text linguistics: aboutness analysis, intertextuality and the cognitive act of classifying documents. J. Doc. 43(1986) No. 2, p. 84-113.

(16) Lufkin, R.C.: Determination and analysis of some parameters affecting the subject indexing process. Cambridge, Mass.: Massachusetts Institute of Technology, Department of Electrical Engineering, Electronic Systems Laboratory 1968. 47 p.

(17) Himanka, J., Kautto, V.: Translation of the Finnish abridged edition of UDC into General Finnish Subject Headings. Int. Class. 19(1992) No. 3, p. 131-134, 139.

Prof.Vesa Kautto, Department of Library and Information Science, University of Oulu, Linnanmaa, SF-90570 Oulu, Finland