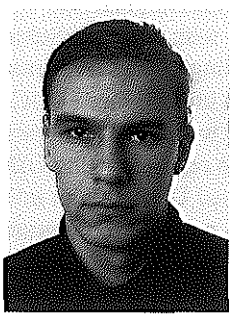
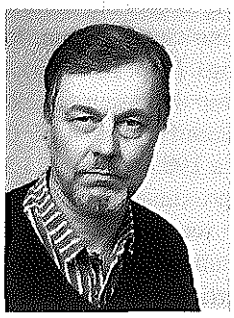
Janne Himanka; Vesa Kautto
Department of Library and Information Science
Oulu University, Finland

Translation of the Finnish Abridged Edition of UDC into General Finnish Subject Headings





Himanka, J. & Kautto, V.: **Translation of the Finnish abridged edition of UDC into General Finnish Subject Headings.**
Int.Classif.19(1992)No.3, p. 131-134, 11 refs.
A method to convert assigned UDC numbers into subject headings is examined. The method consists of two phases. In the first phase, a dictionary is created that maps UDC numbers to subject headings, and in the second one this dictionary is mechanically applied to convert the bibliographic databases. The first phase is tested with 15 different parts of the UDC, with translation success and problems recorded. The problems of each part are scrutinized, and three main types of UDC classes are described: the synthetic, the shallow enumerative and the deep enumerative. The time spent in the first phase is found tolerable, and the initial question of translatability of the whole UDC is considered in the light of the results. (Authors)

1. Introduction

In June 1990 the Department of Library and Information Science in Oulu University started a project to investigate the possibility of translating existing UDC numbers into subject headings of the General Finnish Thesaurus, the standard indexing tool of the National Bibliography of Finland. The rationale behind this goal was the fact that UDC has for decades been the most important indexing tool in Finnish research libraries, and yet recent surveys show that users shun it and use word searching in online catalogues whenever possible. Subject headings have been used in indexing only recently, and not all research libraries use them even today. If it was possible to convert a substantial part of that existing classification data, it would remarkably enhance the searching facilities of online library catalogues. It would offer a uniform word search mechanism for the whole collection instead of the present need to use UDC for older documents.

A complete description of the project can be found in (1), which also contains theoretical considerations on the relationship between classification and thesauri and an overview of the current search-enhancing mechanisms of OPACs. This article concentrates on the results of the experimental parts of the project.

The translation should be a low-cost process, we planned thatthis could be achieved by dividing the work into two distinct phases: first, a dictionary is to be compiled manually that maps UDC class numbers to subject headings, and second, this dictionary is to be applied by a computer to convert the existing databases. The manual effort is thus restricted to the dictionary-forming phase. The result shows, however, that while a totally automatic second phase may be possible, it might not give desirable results.

The work consisted of forming part of the dictionary, evaluating the quality of results and the time that it would take to form a dictionary of the whole UDC (and whether it is possible at all). The index of the Finnish Abridged Edition of the UDC is available online in the KDOK database UDKH. It was used as the source of UDC numbers. UDKH does not have definitions, only index entries, so definitions were checked from the paper version of the UDC.

It was not possible to test the dictionary produced except in a minimal setting, because it covered only about 5% of the whole UDC, and the conversion program was not available. The implementation of the program deserves consideration, and it would not have been meaningful to test the dictionary without the program.

2. Classification and Indexing in Finland's Research Libraries

In the joint statistics of the Finnish research libraries there are 17 university libraries and 3 art university libraries. The statistics includes also 27 major special libraries. There are several small libraries that do not belong to the statistics. The university and art university libraries cover 85% of the collections of the libraries mentioned in the statistics, and also in many other respects they represent more than 80% of the resources and outputs of the Finnish research libraries.

The use of UDC in Finland began in the 1950s. Among the first users were the library of the Helsinki University of Technology, the State Technical Research Centre, Helsinki School of Economics Library and the Oulu University library. The first abridged Finnish UDC edition was published in 1956 by Tietopalveluseura, followed by the second edition in 1967 and the third edition in 1975. A revised version of the third edition was published in 1983 and an index to it in 1985. (3, p. 19-20). This revised edition comprises about 17 000 class numbers.

In the year 1977 a survey was made about bibliographical description. The survey was directed towards the large scientific libraries. There were 30 responses, and of these 30 libraries 17 used UDC, 10 exclusively. Seven libraries had their own classification, and no more than 5 libraries

used subject headings, either exclusively or with other indexing tools. Four libraries used the National Library of Medicine classification, together with Medical Subject Headings. (4, p. 142-143). According to the survey made by Tietopalveluseura in 1982, 123 libraries used UDC as their classification. 44% of these were private libraries (5). In a 1989 survey 311 libraries reported using UDC, but approximately half of them used UDC only in information retrieval (4). The Finnish national bibliography uses shallow UDC, as does the database of Finnish periodical articles, KATI.

In any case, UDC has had a very important role in the Finnish research libraries' bibliographical description. Arja-Riitta Haarala has outlined this history in (7). In the course of decades great systematic catalogues have been developed with UDC. All of the Finnish university libraries have adopted or are in the process of adopting the same online public access catalogue, VTLS (Virginia Tech Library System). It was modified for Finland to enable searches with exact and truncated UDC class number. VTLS, Inc. recommended converting UDC numbers subject headings, however, and this recommendation was one of the roots of our project (8, p. 21). The main motivation for this is the negligible use of UDC in the literature searching process. This has been verified by analyzing the transaction logs of the bibliographic databases. UDC's portion of the search elements was about 3-5% , while subject headings ranked between 26% and 31 % in those databases that had been indexed with them. This supports the idea of translating UDC into subject headings (9).

There is no comprehensive data on the use of the General Finnish Thesaurus. National bibliography indexes a major part of its monographs with it, and the Lapland Univ. Library uses it for indexing all literature.

3. The General Finnish Thesaurus

The General Finnish Thesaurus was composed in the years 1981-1988. The current edition is still a pre-print; the first edition is being worked on and it should soon be available. The thesaurus is maintained by Helsinki University Library, and it thus has quite a solid standing.

GFT in its current form has 10880 subject heading entries and 1934 SEE-references. It consists of two parts: an alphabetical and a systematic one. The alphabetical part dominates the thesaurus; it covers about 4/5 of the pages. The entries are in a typical thesaurus-like arrangement. The alphabetical part contains embedded hierarchical relations, as well as related term-references. The hierarchies are quite shallow, and not ubiquitous, so GFT is somewhere between subject heading lists and true thesauri. Only immediate predecessors and successors in the hierarchy are shown. Polyhierarchies exist, but they are more of an exception than a rule.

The systematic part of GFT is divided into 61 groups, roughly in the two- or three-number UDC specificity level. The groups have no further divisions. The systematic part is designed to give an overview of the field's terminology. A given term can belong in several groups. The group codes are also visible in the alphabetical part, to give the user some knowledge of the term's intended scope (2).

In the beginning of GFT there are some brief recommendations on use. Suffice it to mention that chains of terms can be used, and a colon serves as the relational operator. The terms in a chain can be equal or one term can explain another. If the terms are equal, all of them will be searchable.

4. The Mechanics of Translation

About fifteen classes of the UDC were chosen to be translated. The most important selection criterion was that the classes should represent the heterogeneity of the UDC as well as possible. The selected classes were:

+ .00 Common auxiliaries of point of view
+ -05 Common auxiliaries of person
+ 00 Generalities
+ 159.9 Psychology
+ 23 Christian Dogmatics
+ 316 Sociology
+ 502/504 Environmental sciences
+ 539 Physical nature of matter
+ 677 Textile Industry
+ 73/76 Pictorial arts
+ 800/801 Linguistics, Philology
+ 82 Literature

The edition of UDC used as a source was the Finnish Abridged edition, printed in 1983, and the revisions that have been made after that. The actual modus operandi was to insert the translations to records that were retrieved from the online UDC index, UDKH. The translation thus produces a record consisting of three different fields: the UDC class numbers, the subject heading translations and the UDC index terms. This brought to light several inconsistencies between the index and the tables. The Finnish UDC index is undergoing revision, but at the time of the translation the new version was not available.

A typical result of the translation would look like this:

UDK: 159.925.8
AS: gestures
AS: expressions
HA: gesticulatory communication
HA: mimicry

Here UDK: means the UDC number, AS: means subject heading (the translation) and HA: is the UDC index term plus its scope note. The example is translated from Finnish and may deviate from the corresponding English entry.

5. The Three Types of Classes

The translation of the several UDC broad classes revealed three main types of different classes (from a translation point of view). They are called here synthetic, shallow enumerative and deep enumerative types. The types are meant to capture some of those characteristics of classes that affect the translation process.

5.1 Synthetic

By synthetic classes we mean those broad classes whose member classes often occur in combination with other classes in a document classification. Strictly speaking, the means of combination should be the colon operator of the UDC. Often the same effect is achieved with several separate UDC numbers, however.

Typical synthetic classes in our sample were 73/76, 800/801 and 82. The hierarchy in these classes is shallow, and resembles the intra-facet hierarchies of Colon Classification. These classes show quite a few traces of facet analysis.

The translation of synthetic classes was smooth and unproblematic for the most part. The ease of translation will probably be countered by difficulty of parsing. The parsing of complex synthetic class numbers is not as easy as parsing the normal main class—common auxiliary—special auxiliary scheme, but it should not be too difficult. For example, the 82 class (Literature) uses common auxiliaries extensively, and this should not pose a problem since they have to be dealt with anyway. The special auxiliaries of 82, however, are indeed special. The parser will have to know that special auxiliaries can be obtained under 800, if they are not found under 82 itself. Similarly, under 803/809 the special auxiliaries are in fact subdivisions of 800 and 801. Some general device for dealing with this kind of auxiliaries will be necessary. A similar phenomenon in enumerative classes is the "Subdivided as" -mechanism.

Despite the perceived simplicity of translating synthetic classes they demand perhaps more thought than the enumerative classes, because the translator should anticipate the many roles in which the class number can be used. When a class number is used in an enumerative mode, its role is always the same: it designates the place in the UDC hierarchy that a document occupies. In contrast, synthetic class numbers - and even more the common auxiliaries - can have a variety of different roles: they can be objects, agents, attributes and so on.

5.2 Shallow Enumerative

Enumerative classes with shallow hierarchic structure were usually rather straightforward to translate. For example, in the 23 class (Christian dogmatics) only 12.3 % of the class numbers remained without a subject heading counterpart. Translating these classes is hit-or-miss; usually one cannot paraphrase if precise equivalents do not exist. Another example of this class type is 159.1, Psychology.

Classes of this type frequently employ an unfortunate device of cutting branches that are too deep for an abridged edition of UDC. The device works so that the cut branches are not abandoned altogether but are collected into a single class number. Usually the concepts are siblings from the immediately subordinate hierarchy level of the deeper edition. While this may be quick and easy for the compiler of the abridged edition, it is an abomination to one who tries to translate classes into subject headings. An example of this is 316.343.65. The definition of this class in the Finnish abridged edition is Intermediate social strata. Middle class. Intelligentsia. Managers. White collar workers. Civil servants. In the English full edition this class is defined only by the first phrase, and the other concepts are children of this class. There are a few different solutions for handling these cases. First, one can dutifully repeat all the different concepts as subject headings. This is what Riesthuis and Bliedung (10) have done in their experimental thesaurification of the demography class of the UDC. This leads to problems, however. The unwritten rules for classification and indexing are different: a composite class number that enumerates several concepts well deserving their own class numbers can be used to describe a document that only deals with one of these aspects. On the other hand, we cannot describe a document with 5 unrelated subject headings, even if these subject headings were generated from the above mentioned composite UDC class number. This would be considered misleading.

Another method for handling these flattened hierarchies would be inventing a higher-level concept that encompasses all the terms that are enumerated in one UDC class number. This is the ideal solution, but it is rarely possible. Presumably the compilers of the abridged edition would have used a single concept had they found one. Also, one might consider checking the full edition, since it might actually be easier to translate the original structure.

When the target of the translation is a controlled vocabulary, one's ambitions are often checked by inadequacies in the thesaurus, regardless of its quality. UDC differs from GFT so extensively in structure that cul-de-sacs in translating are quite common. The translator can decide to apply the two methods mentioned above, but often this becomes impossible, if there is no way to translate a class number that has adopted five full-edition subclasses, and the entry has to be skipped over. If we accept the idea of human intervention in the conversion process, it would be possible to make the program produce warnings when it encounters a class that has been flagged as problematic in the dictionary-forming phase.

5.3 Deep Enumerative

From the enumerative parts with deeper hierarchic structure two distinct types of classes emerged. An example of the first one is the textile industry class 677. Special auxiliaries form a substantial part of 677 class. They are used almost exclusively in combination with the 677 class, however, and the borderline between main classes and special auxiliaries seems rather arbitrary in this case.

The abridged edition often uses an "Other ..."-mechanism in cases where further subdivision is unwanted and a reasonable concept covering all the more specific classes does not exist. This was a particularly painful construct for the translator, since you cannot index a book with phrases like "Other fabric-forming machines". These

classes usually were on the most specific level of the hierarchy, and were left out of the translation. If the class was subdivided further (i.e. was used as a "subheading"), translation was sometimes possible by translating the children in place of the parent.

The main characteristic of this type of class is that translation becomes more unsuccessful in a linear fashion as the classes become more specific. 43 % of the classes in 677 could not be translated, if every class was considered. However, if we only translated 40% of the class numbers, choosing the most general, the percentage of the untranslatable classes was only 8. This suggests that it could be reasonable to leave the most specific class numbers out of the translation if it is apparent that the translation result deteriorates linearly as the specificity increases.

An example of the other kind of deep enumerative class is 539, The physical nature of matter. Here the deterioration of translation success as a function of specificity was not linear. There were untranslatable classes everywhere in the hierarchy, indeed, 45 % of the classes were untranslatable. When translation was restricted to 1/3 of the classes, leaving out the most specific 2/3, the amount of untranslatable classes was still 25 %. This was the only class in the sample where a satisfactory result could not be achieved by any means. The untranslatable classes in 539 were very specific physical terms, which hardly belong to a general thesaurus of less than 20 000 entries.

It would seem that classes like 677 can be translated with moderate success if it is acceptable that the most specific classes are stripped. Translation seems impossible only when the untranslatable classes cannot be eliminated systematically.

6. Conclusion

Of all the classes in the sample, 16.6% were untranslatable. Record was kept during the translation about the translation success. This record shows that according to the translator 73% of the translations were on the same specificity level as the original class number. In a large perspective, the most problematic parts of the UDC to the translator are the specific enumerative classes. This depends on the target vocabulary, and may not be a problem to one who uses e.g. LCSH as the controlled vocabulary. In a smaller scale, the most frequent obstacles were the flattened hierarchies that are common in abridged editions. After doing some comparisons with the full edition, it indeed seems that this problem has one of its roots in the way the abridged edition was compiled. This, and more generally the difficulty of treating class definitions that are unsuitable for word searching, is already known to index compilers (11).

The picture we got of the UDC is quite a colourful one. UDC's structure varies remarkably, obviously reflecting the historical phases the system has gone through. The older parts, showing the Donker Duyvis influence, are fashioned according to Dewey's classification and possess a clear hierarchical model, while newer classes emphasise synthesis and exhibit traits of facet analysis. These shifts have their consequences for an attempt to render the system into another vocabulary. The synthetic, 'faceted' parts can be translated without much effort, but the chains they produce may seem cryptic to the user; the 'enumerated' parts, contain specific and obsolete material that is unlikely to be translated well into a modern controlled vocabulary, but the resulting translations are more intuitive, since there is no need to combine subject headings into chains.

Results from translating the deep enumerative classes 677 and 539 seem to give a negative answer to the question that was posed in the beginning. It does not look feasible to attempt a translation of the whole abridged UDC. However, as the untranslatable parts of the UDC are few and possibly identifiable, it does not seem unreasonable to consider a partial translation, concentrating on those parts that are supposed to give good results. The required amount of work does not seem to be prohibitive; judging from this project, translating all possible classes would take about one work year.

There are no signs of immediate adoption of our scheme in Finland. What is under way is that the new UDC index will build upon the terminology of GFT where it is applicable. Another trend, related to this, is linking the UDC index terms to the class numbers, so that people using VTLS can make use of UDC's structure, even if they are searching with words. As this was also the incentive of our work, one can conclude with Catullus's words: "O dulces comitum valete coetus,/longe quos simul a domo profectos/ diversae varie viae reportant."[1]

Note

1 In English: *Farewell, dear bands of travellers, who started together from your far-away home, and whom different ways from different directions are bringing back again.* (Catullus, XLVI,9-11)

References

(1) Himanka, J.:UDK: n suomalaisen lyhennetyn laitoksen kääntäminen Yleisen suomalaisen asiasanaston termeiksi (Translating the Finnish abridged edition of UDC into General Finnish Subject Headings). MSoc.Sc. thesis in preparation. Oulu 1992.

(2) Yleinen suomalainen asiasanasto (The General Finnish Thesaurus). Edited by R. Maajärvi, E. Kärki and S. Ojonen. Helsinki: Kirjastopalvelu 1988.

(3) Huiku, L.: UDK kirjastoluetteloissa (The use of UDC in library catalogues in Finland). Tampere: Tampereen teknillinen korkeakoulu 1990.

(4) Kautto, V.: Suomen suurimpien tutkimuskirjastojen luokitus- ja indeksointijärjestelmät ja atk-systeemien tiedonhakuun liittyvät toivomukset (The classification and indexing systems of the major Finnish research libraries and the requirements concerning IR with automated systems). Signum 11(1978) No. 7, p. 142-146.

(5) Kuuteri, R.: UDK-järjestelmää käyttävät kirjastot Suomessa (The Finnish UDC-using libraries). Signum 15(1982) No. 10, p. 232.

Continued on page 139