Eugeniusz Scibor, Joanna Tomasik-Beck
Institute for Scientific, Technical and Economic Information, Warsaw

On the Establishment of Concordances between Indexing Languages of Universal or Interdisciplinary Scope (Polish Experiences)

Prof. Eugeniusz Scibor, graduated from the Faculty of Library Science of the Warsaw University. Doctoral degree in library and information science, the degree of doctor habilitus (assistant professor) in foundations of information science. His interests are: theoretical foundations of indexing languages especially of classification systems), compatibility of indexing languages.

Dr. Joanna Tomasik-Beck, graduated from the Faculty of Polish Philology of the Warsaw University. Doctoral degree in terminolgy science. Her interest are: indexing language, information retrival systems, terminology sciene.

Scibor, E., Tomasik-Beck, J.: **On the Establishment of concordances between indexing languages having a universal or interdisciplinarg scope** (Polish experience)
Knowl.Org. 21(1994)No.4, p.203-212, 10 refs.
Reports on investigations conducted at the Institute for Scientific, Technical and Economic Information (ISTEI) in Warsaw (Poland) in 1992-1993. These investigations aimed at a comparative analysis of four indexing languages (ILs) developed and/or used in Poland and at the establishment of concordances between them. These were the following ILs: Polish Thematic Classification (PTC) descriptor language based on the Thesaurus of Common Topics (TCT), Universal Decimal Classification (UDC), Subject-Heading Language of the National Library in Warsaw (SHL). When establishing concordances the PTC was chosen as the master language whereas the three other ILs served as target languages. The research in question comprised: a) pilot investigations; b) main investigations consisting in the elaboration of the Concordance Dictionary of Indexing Languages (CDIL). The pilot investigations comprised three approaches: a) use of a random sample of 144 PTC headings to which the eqiuvalent lexical units of the target ILs were assigned; b) generation of alphabetical comparison matrices (MI) enumerating the lexical units of the ILs under investigation belonging to four selected subject-fields; c) establishment of concordance tables between the PTC and the three target ILs within the same four subject-fields. The elaboration of the CDIL consisted in the assignment of the (more or less) equivalent lexical units of the three target ILs to all 1330 headings comprised in the PTC main table. The coincidence rate of the ILs under comparison was computed in the framework of the pilot investigations as well as when analysing the results of the elaboration of the CDIL. The computed coincidence rate was very low when only the full (exact) equivalence of the lexical units belonging to the ILs under investigation was taken into consideration but it considerably increased when also the partial equivalence was taken into account. (Authors)

1. Preliminary considerations. A short description of the indexing languages taken into account when establishing concordances

In Poland - as in other countries - various indexing languages (ILs) are used. Most of them were elaborated independently and are incompatible with each other. So the problem of comparing the vocabularies of these ILs and establishing concordances between them was considered important.

In 1993 the so-called "Concordance Dictionary of Indexing Languages" (CDIL) was elaborated at the Institute for Scientific, Technical and Economic Information (ISTEI) in Warsaw. The elaboration of the CDIL was made possible by a govemmental grant which had been awarded by the Polish Committee for Scientific Research. The CDIL is a table of concordances between four ILs having a universal or interdisciplinary scope, elaborated and/or used in Poland, namely between the Polish Thematic Classification (PTC) 1), the descriptor language based on the Thesaurus of Common Topics (TCT), hereafter simply called TCT, the Universal Decimal Classification (UDC) and the Subject-Heading Language of the National Library in Warsaw (SHL).

It is a well known fact that the elaboration of concordances requires the discerning of one master language and of one or several target languages. The PTC was chosen as master language for the following reasons: 1) it has a universal scope; 2) its structure is hierarchical; 3) the limited bulk of its vocabulary ensured the feasiblity of the project. The PTC is a shallow, in the main monohierarchical (uni-dimensional) classification which encompasses practically all fields of knowledge, arts and branches of economics in which Polish information establishments (i.e. information and documentation centres, libraries, etc.) are likely to be interested. The depth of the PTC main table is limited to three levels of hierarchy. A uniform centesimal notation system has been adopted.

At present the main table of the PTC contains 1330 headings (entries). As in the majority of other classification systems each PTC heading consists of the appropriate PTC number and of the wording which explains the meaning of the number. Some of the PTC headings are annotated. The annotations have the form of so-called 'contents descriptions': the expressions appearing in the 'contents descriptions' could be easily converted into separate PTC headings which would constitute the fourth level of the hierarchy.

On the first level of hierarchy eighty PTC main classes appear which are ordered within the following four thematic groups:

- social sciences (including the humanities);
- natural sciences (mathematics, physics, chemistry, earth and biological sciences);
- applied sciences (medicine, technology, etc.) and branches of economics;
- general and complex (interdisciplinary) problems (information science, science of science, statistics, environmental protection, etc.)

Apart from the main table there exist also three series of auxiliaries, namely the so-called standard subdivisions (similar to some extent to the common auxiliaries concerning the point-of-view in the UDC) as well as geographical and language ones. As in other classification systems these auxiliaries are used to build up the compound PTC numbers which denote complex concepts. Of course, there exist also alphabetical indexes to the main table and to the auxiliaries. The newest (third) version of the PTC has not been published yet but has been converted into a machine-readable form using the Micro CDS/ISIS 3.0. software and is available on diskettes and/or in the form of computer printouts.

The PTC has been developed at ISTEI since 1976. It is to be used first of all to determine the scopes of the document collections, sub-systems and services which are created and developed in the framework of large (first of all interdisciplinary) information systems. The PTC can also be used to determine the scopes of branch indexing languages (branch classification systems, descriptor languages based on branch thesauri, etc.). The classification system in question can also serve as a tool for arranging the contents of information publications (abstracting journals, guidebooks, etc.). Incidentally - if the need arises - the PTC can also be used (preferably together with other indexing languages, e.g. with keyword systems) for information retrieval (2). Another possible application of the PTC consists in serving as a starting-point for the elaboration of branch classification systems. The methodology of constructing such systems - which would constitute enlargements of the PTC within given fields of knowledge and/or branches of economics - was elaborated at the ISTEI in 1986 (3).

As the first of the three target languages used for the construction of the CDIL let us mention the TCT. As a rule the TCT cannot be used as the only indexing tool in any information system. Instead, it can be used for indexing documents and queries in practically every information system together with a thesaurus containing the essential vocabulary of that system. The TCT serves first of all to provide the designers of thesauri destined for various fields of knowledge and/or branches of economics with a ready standardized vocabulary having an interdisciplinary, "cross-sectional" character; this function of the TCT should be considered as primordial.

The scope of the TCT - which was described at great length in (4) - encompasses:

- general problems (generalities),
- social and economic problems,
- problems connected with natural sciences (first of all environmental pollution and protection),
- technical problems.

Moreover, the thesaurus in question contains some names of persons, terms denoting fields of human activities (fields of knowledge, branches of economics, etc.), names of places (including geographical names) and indications connected with time.

As in other thesauri the vocabulary of the TCT consists of descriptors and non-descriptors (forbidden terms). The total number of descriptors is about 3000 and that of non-descriptors about 400. Also as in the majority of other thesauri, in the TCT two kinds of relationships were introduced, namely the relationship of equivalence between descriptors and non-descriptors and the paradigmatic (i.e. hierarchical and associative) relationships between particular descriptors.

The newest (third) version of the TCT was elaborated at ISTEI in 1987. It consists of the introduction, the systematic part, and the alphabetico-hierarchical part. The systematic part has a fully faceted structure; the vocabulary of the TCT was arranged with twelve subject categories, such as Branches of science and of economics, Legal instruments and regulations, Scientific theories and doctrines, Persons, Organizations and institutions, etc. Each category (except category XII, Time) is divided into several facets, e.g. category IV, Persons, comprises two facets: IV.1. Individuals and IV.2. Groups of persons. The alphabetico-hierarchical part is the main part of the TCT since it contains the full information on all thesaurus terms (descriptors and non-descriptors) (5).

At present a computer-readable version of the TCT does not as yet exist, so this thesaurus is available only in the form of a typescript (not yet published). It is planned to enlarge the TCT and convert it into the Polish Macrothesaurus which - of course - will be maintained in machine-readable form.

The UDC - which is very widely used in Poland - was chosen as the second target language. This classification system is well known all over the world so there is no need to describe it in this article. When elaborating the CDIL the UDC edition mainly was the "Abridged edition for the Polish network of scientific information" (6). This edition has some features of a special subject UDC edition (i.e. the wide use of compound and complex UDC numbers); it was not published yet when the CDIL was elaborated, but the authors of the project made use of its manuscript.

The SHL, which played the role of the third target language, has a quasi-natural (paranatural) language vocabulary and a positional grammar; by the 'quasi-natural' (or 'paranatural') language vocabulary the use of natural language expressions as the lexical units of an indexing language is meant, whereas by the "positional grammar" a fixed word order in a sentence is meant - each translocation of a given word within a sentence changes the meaning of the sentence as a whole. As in the case of other subject-

heading languages the vocabulary of the SHL consists of two kinds of lexical units, namely of the subject-headings in the proper sense (we can call them also "leading subject-headings") and of subject *sub*headings. The (leading) subject-heading constitutes the main, obligatory element of a sentence (i.e. entry) formulated in a subject-heading language and is always placed at the beginning of such a sentence. A subject subheading usually indicates a formulation of a secondary feature (aspect) of the subject of a document. A subject subheading can also express the form of a document as a literary product or as a kind of publication; such subheadings are called 'formal subheadings'. The sentence formulated in a subject-heading language (we can call such sentence 'subject-heading entry' in order to distinguish it from a subject-heading as a lexical unit) consists of one (always only one) subject-heading and of one or a few subject subheadings; sometimes it comprises only the sole subject-heading without any subheadings attached to it. In the subject-headings languages which have been developed in Poland the sign '-' (dash) separates the (leading) subject-heading from the subject subheading (or subject subheadings) which is (or are) placed within the same subject-heading entry; the same sign is used to separate particular subject subheadings occurring within the same entry.

The vocabulary of the SHL is contained in the *Dictionary of the Subject-Heading Language of the National Library*. This *Dictionary* comprises two parts which can be considered as main ones. The first main part is called "Subject-headings and references" and contains the full list of subject-headings as well as the so-called 'rejected terms' which play the same role in the said *Dictionary* as non-descriptors in thesauri. This part of the *Dictionary* is equipped with a set of references which express several kinds of relationships occurring between the terms contained in the *Dictionary*, namely the relationship of retrieval equivalence (retrieval correspondence) between rejected terms and subject-headings (and vice versa) as well as the hierarchical and associative relationships between particular subject-headings. All these kinds of relationships are also denoted in the majority of thesauri, though the references expressing these relationships usually take a different shape in the dictionaries of the subject-heading languages (such dictionaries are traditionally called 'alphabetical lists of subject-headings') than in the thesauri. In the *Dictionary* in question appears also a specific type of reference which seldom if ever occurs in thesauri. These are the so-called multiple references; such references do not refer the user to a concrete lexical unit (or to a few such units), but only indicate the direction of searching for the appropriate subject-headings, e.g.:

Physiology
see also the relevant subject-headings with the subheading - physiology, e.g. Plants - physiology; Sport - physiology

The second main part of the "Dictionary" (which is, however, much smaller than the first one) enumerates the subject subheadings that are divided into a few categories according to the kinds of subject-headings with which they are used when formulating subject-heading entries (subject subheadings used with general subject-headings, subject subheadings used with geographical subject-headings, etc.). There are some subject-subheadings which can be used only with one concrete subject-heading, e.g. the subheading "rendering accessible" can be used only with the subject-heading 'Library holdings'.

When elaborating the CDIL use was made of the first edition of the *Dictionary* published in 1989 (7). Only after completing the CDIL the second, improved and enlarged edition of the "Dictionary" in question (8) was published. It is a pity that the compilers of the CDIL could not use this second edition, as it seems much better than the first one; some mistakes in expressing the relationships between subject-headings have been eliminated and the presentation of relationships is now very similar to that used in the majority of thesauri.

2. Pilot investigations

Before attacking the task of constructing the CDIL its compilers assumed that the vocabularies of the four above-mentioned ILs: 1) overlap with each other; 2) are ordered by means of hierarchical structures and other kinds of relationships in a manner that permits the elaboration of the CDIL, i.e. ensures its feasibility.

In order to confirm the truth of these assumptions pilot studies were made. An additional aim of the experiment - which took place in 1992 and was conducted by Joanna Tomasik-Beck - consisted in the computation of the verbal coincidence rate of the ILs under investigation (9).

When executing the said pilot studies two research methods were used. The first method consisted in the use of a random sample of PTC headings representing the whole scope of this classification, i.e. practically the whole universe of knowledge. Making use of the random method the author of the studies selected 144 PTC headings; this set of headings constituted about 10,8% of the total number of headings contained in the main table of the PTC. Then to each PTC heading thus selected she assigned the equivalent units pertaining to the three above-mentioned target languages. The process of assignment was very difficult and necessitated the utilization of all possibilities inherent in each of these ILs. Thus in the case of the TCT the assignment of two or more TCT descriptors to one PTC heading was very often necessary; this is shown in Table 1.[2].

PTC heading	Equivalent TCT descriptors
68.02.00 History of medicine	I.1 History I.1 Medical sciences

Table 1: Example of the assignment of two descriptors from the TCT to one PTC Heading.

As far as the equivalence between the PTC and the UDC is concerned, simple, compound and complex UDC numbers were assigned to the PTC headings contained in the random sample. This is shown in Table 2.

PTC headings	Equivalent UDC numbers
23.30.00 Theory of groups	512.54 (simple number)
29.12.00 Present-day research and exploratory expeditions	910.4"312" (compound number)
14.09.19 Psychology of mass communication	316.77:659.31.001:316.27 (complex number)

Table 2: Examples illustrating the assignment of various kinds of UDC numbers to PTC headings.

In the case of the PTC -> SHL equivalence (equivalence to the PTC headings appearing in the random sample) either subject-heading entries were assigned which consist only of the subject-heading alone or entries comprising the (leading) subject-heading as well as one subject subheading or two subheadings (see chapter 1. of this article). Table 3. below illustrates the three above-mentioned situations.

PTC headings	Equivalent SHL subject-heading entries
10.23.00 Social psychology	Social psychology
74.09.00 Theory of trade	Trade - theory
47.19.31 Industry of asphalt and modified asphalt products	Asphalt - products - industry

Table 3: Examples of the assignment of various kinds of SHL subject-heading entries to PTC headings.

Although various possibilities characteristic of each of the target languages (including the possibilities resulting from the very thorough application of the grammar of these ILs) were utilized, the exact equivalents in all three languages could be assigned only to 19,4% of the PTC headings contained in the random sample. In this connection broader, narrower or otherwise semantically related lexical units were assigned to the PTC headings when fully equivalent units were lacking in the target languages. This procedure resulted in the increase of the percentage of the PTC headings having (more or less) equivalent lexical units in all three target ILs to 81,25%.

Next the number and percentage of the lexical units pertaining to particular target ILs which had equivalent PTC headings within the random sample were computed. The results obtained are shown in Table 4.

Target language	Lexical units having more or less exact equivalents in the random sample of PTC headings	
	Number	% of the potential equivalencies between the PTC and the target ILs (100% = 144)
TCT	121	84,0 %
UDC	136	94,4%
SHL	136	94,4%

Table 4: Results of assigning the lexical units of the target ILs to the PTC headings comprised in the random sample.

To 16% of the PTC headings contained in the random sample no equivalent TCT units could be assigned. In the case of the UDC and the SHL the percentage of unsuccessful assignments amounted to 4,2%.

The second method used when carrying out the pilot investigations consisted in the establishment of comparison and compatibility matrices according to the method proposed by I. Dahlberg in (10).

To begin with four PTC main classes were selected which belonged to four different subject areas represented in this classification system (1. Humanities and social sciences; 2. Natural sciences; 3. Applied sciences; 4. Technology and various branches of industry). These classes were the following:

08.00.00 State and law
24.00.00 Physics
41.00.00 Computer science
47.00.00 Mineral industry

This choice of classes resulted from the assumption that different kinds of problems may arise within each of the above-mentioned subject areas when establishing the concordances between ILs. It was assumed that the solution of these problems would facilitate the elaboration of the CDIL.

At first for each of the subject-fields represented by the four above-mentioned classes a separate list was established containing the lexical units belonging to the given subject-field and occurring in the ILs under investigation (PTC, TCT, UDC, SHL). On the basis of these lists, four alphabetical comparison matrices (M1) were generated in which - in alphabetical order - all terms representative of the investigated subject-fields and appearing at least in one of the ILs under comparison were enumerated. Next, to each expression occurring in the list numbers (codes) and wordings (verbal formulations) used in the ILs under investigation were assigned[3]. Through an analysis of comparison matrices thus constructed and encompassing the four above-mentioned subject-fields it was possible to ascertain to what degree the vocabulary of each of the ILs under comparison overlaps in the lists elaborated for particular subject-fields.

It was found that - within all four subject-fields - the UDC had the richest vocabulary; UDC numbers could be assigned to 94,8% of the expressions concerning the mineral industry; the corresponding figures relating to state and law, physics and computer science amounted to 78,4%, 76,8% and 70,2% respectively. In the case the SHL the vocabularies relating to state and law (62,5%) and physics (54,4%) were well represented, whereas in the TCT there was a good representation of the terminology concerning state and law (58,2%). The smallest number of terms was registered in the case of the PTC, which is characterized by a high degree of the generality of the vocabulary.

On the basis of the comparison matrices thus elaborated the occurence of the expressions of the generated master

language in the PTC (language a), TCT (language b), UDC (language c) and SHL (language d) within particular subject-fields was investigated.

The following fifteen situations reflecting the appearance of the given expression in the languages (or in the language) were noted:

1. a,b,c,d	2. a,b,c	3. a,b,d	4. a,c,d	5. b,c,d
6. a,b	7. a,c	8. a,d	9. b,c	10. b,d
11. c,d	12. a	13. b	14. c	15. d

The results obtained are presented in Table 5.

Serial number	Appearance of the lexical units in the ILs under investigation	Number of lexical units			
		State and law	Computer science	Physics	Mineral industry
1.	a b c d	17	8	4	-
2.	a b	2	3	-	1
3.	a c	-	1	8	5
4.	a d	1	2	3	-
5.	b c	25	1	-	2
6.	b d	14	3	4	-
7.	c d	51	5	66	34
8.	a b c	3	2	-	2
9.	a b d	1	2	-	-
10.	a c d	1	4	6	-
11.	b c d	48	8	1	2
12.	a	1	6	3	3
13.	b	23	7	-	-
14.	c	31	31	96	47
15.	d	14	-	44	1

a = PTC
b = TCT
c = UDC
d = SKL

Table 5: Occurrence of the PTC, TCT, UDC and SHL lexical units belonging to the four selected subject-fields

In virtue of the data contained in Table 5 the following compatibilities were computed within the subject-fields: state and law, physics, computer science, mineral industry: the synchronous compatibility of the a, b, c, d languages (abcd) as well as the compatibility of a with b (ab), a with c (ac), a with d (ad), b with c (bc), b with d (bd) and c with d. The computation was done using the following formula:

$$(1) \qquad C_{abcd} = \frac{\sum_{i=1}^{n} (a_i b_i c_i d_i)}{n}$$

It was ascertained that the verbal coincidence of the ILs under investigation in the fields of mineral industry, computer science, physics and state and law was very low. This was shown in Table 6.

Subject-fields	Languages						
	a b c d	a b	a c	a d	b c	b d	c d
Mineral industry	0	0,2	0,074	0	0,065	0,048	0,387
Computer science	0,028	0,319	0,205	0,364	0,253	0,467	0,415
Physics	0,056	0,138	0,096	0,094	0,027	0,07	0,332
State and law	0,118	0,158	0,11	0,123	0,431	0,4	0,568

a = PCT
b = TCT
c = UDC
d = SHL

Table 6: Verbal coincidence of the TC, TCT, UDC, and SHL lecial unites concerning the four selected subject-fields.

In the case of the mineral industry the verbal coincidencerate amounted to 0, because no expression occurring in all four ILs was noted. As far as the other subject-fields under investigation are concerned the verbal coincidence rate amounted to 0,028 (computer science), 0,056 (physics) and 0,118 (state and law) respectively. It appeared that the UDC and the SHL were those ILs which were the most compatible with each other, especially in the field of state and law where the coincidence rate amounted 0,568. A high coincidence rate was noted between the TCT and the SHL in the fields of computer science (0,467) and state and law (0,4).

At the next stage of the pilot investigations the compatibility matrix (M2) for the subject-field 'mineral industry' was elaborated in which the alphabetical arrangement was abandoned and all terms were hierarchically ordered. Thus the master language x came into being. To each lexical unit of this language the corresponding units from all four ILs under investigation were assigned. Different kinds of compatibility were taken into account: full compatibility, assignment of broader and narrower terms, semantic relationship of nonhierarchical nature. Then the degree of the conceptual compatibility of the ILs under investigation was computed: it amounted to 0,11 and was higher than the degree which was computed when constructing the comparison matrix (M1) for the subject-field 'mineral industry'.

It appeared that the generation of the compatibility matrix (M2) - in which all lexical units should be hierarchically ordered - for all subject-fields under investigation would be very time-consuming anf difficult. In many cases the generation of such a matrix would be quite impossible because the same terms appearing in different ILs belong to different hierarchical chains created in these languages. Another difficulty consisted in the fact that the vocabulary of the UDC and the SHL was much more detailed than that of the two other languages under investigation (i.e. the PTC and the TCT).

Afterwards concordance tables were elaborated between the PTC (master language) and the three other ILS under investigation (TCT, UDC, SHL) which played the role of target languages; the concordances covered all four selected subject-fields (state and law, physics, computer science, mineral industry). Then the coincidence rate was computed with the aid of formula (1).

To the PTC headings concerning the four above-mentioned subject-fields the lexical units constituting the full or partial equivalents of these headings were assigned; in some cases broader or narrower terms were assigned. It was found that the coincidence rate of the ILs under investigation had considerably increased; it amounted to 0,96 (computer science), 0,94 (state and law), 0,67 (physics) and 0,58 (mineral industry).

On the basis of the results of the pilot studies the following conclusions were formulated:

I) A comparison of the four ILs under investigation (PTC, TCT, UDC, SHL) permits us to ascertain - as far as the four selected subject-fields are concerned - that the ILs in question differ in the numerical force of the vocabulary, which causes their low coincidence rate.

2) As was expected, the elaboration of the CDIL is only possible if the PTC is chosen as master language.

3) The elaboration of the CDIL will be possible and advisable if different types of equivalence (i.e. the full equivalence and the partial one occurring when broader, narrower or otherwise semantically related terms of the target languages are assigned to lexical units of the master language) are taken into account when assigning the lexical units of the target ILs to the PTC headings. Owing to the application of such a strategy the compatibility of the ILs under comparison considerably increases and can be considered a high one.

3. Elaboration of the CDIL

After the conclusion of the pilot studies the group consisting of three persons (Mrs. Bozenna Klaga, Prof. Eugeniusz Scibor, Dr. Joanna Tomasik-Beck) from the Department of Indexing Languages of ISTEI tackled the job of elaborating the CDIL. Each of these three persons was responsible for the establishment of concordances between the PTC as master language and one of the three target languages; thus E. Sciborwas responsible for establishing concordances between the PTC and the TCT, B. KLaga between the PTC and the UDC, and J. Tomasik-Bec for assigning the SHL subject-headings to the PTC headings. The whole project was executed under the direction of E. Scibor, who assumed the duties of project manager. The elaboration of the concordances was completed towards the end of 1993.

3.1. Methodology and course of the project

As already stated at the beginning of this article the elaboration of the CDIL consisted in the assignment of the more or less equivalent units of the three target languages (TCT, UDC, SHL) to all 1330 headings of the PTC, which had been chosen as the master language.

The construction of the concordances was from the beginning computer-aided. Therefore afore-mentioned Micro CDS/ISIS 3.0 software was used. To each record contained in the main PTC database (comprising the main table of the PTC) a second page was added containing the following fields:

1. Facet code of the TCT
2. Descriptor(s) taken from the TCT
3. TCT -> PTC equivalence symbol
4. UDC number(s)
5. Wording(s) of the UDC number(s)
6. UDC -> PTC equivalence symbol
7. Subject-heading(s) taken from the SHL
8. SHL -> PTC equivalence symbol.

As far as all necessary machine operations were concerned the above-mentioned group of three persons was assisted by Mr. Marek Sulej, a graduate engineer from the Department of Information Technology of ISTEI.

Table 7. shows five one-sign symbols that were used in order to express different kinds of relationships occurring between the PTC headings and the assigned lexical units of the three target ILs.

Symbol	Significance of the symbol
=	Full equivalence (the lexical unit of the target language is completely or almost completely equivalent to the PTC heading)
	Partial equivalence:
<	Broader term (the scope of the lexical unit of the target language is broader than the scope of the PTC heading)
>	Narrower term (the scope of the lexical unit of the target language is narrower than the scope of the PTC heading)
≠	Related term (the lexical unit of the target language is someway semantically related to the PTC heading but the relationship has not a hierarchical character)
0	Zero equivalence (in the target language there is lacking a lexical unit which would be equivalent to the PTC heading)

Table 7: Equivalences ymbols used when establishing the CDIL.

As when carrying out the pilot investigations, broader, narrower or otherwise related lexical units of the target languages were assigned to the PTC headings when the completely equivalent lexical units were lacking in these ILs. In general a broader lexical unit of a target language was assigned to a given PTC heading when in the target language no equivalent lexical units appear on the level of hierarchy corresponding to the PTC hierarchy level on which the given heading was situated, while a semantically related lexical unit did appear on the upper level of the hierarchy. Similarly, a narrower lexical unit of the target language was assigned to a given PTC heading when in the target language no equivalent lexical units appeared on the level of hierarchy corresponding to that PTC hierarchy level to which the given heading belonged, while a semantically related lexical unit did appear on the lower level of hierarchy. Sometimes on the upper or lower level a few broader or narrower lexical units were found; in the case of the assignment of two or more upper-level lexical units to one PTC heading this situation resulted in a kind of polyhierarchy within the structure of the CDIL. There was a certain problem of redundancy of the units to be assigned to a given PTC heading when in a target language the full equivalent was lacking on the level of hierarchy corre-

sponding to the PTC hierarchy level on which the given heading was situated whereas the semantically related units were found on the upper level as well as on the lower one.

Below, some examples are presented to illustrate the method of assigning the lexical units of the target ILs to the given PTC heading when full equivalents were lacking.

PTC 25.23.21 Analysis of organic matters
SHL < Analytical chemistry

PTC 03.01.02 History and the present-day state of the study of religions
TCT > I.1. History ^ Study of religions

PTC 11.35.00 Education in the family. Self-education
SHL > Education in the family

The above examples show that the scopes (extensions) of the given lexical units were first of all taken into account when determining their relationships, e.g. the lexical unit 'Education in the family. Self-education' was considered as a narrower one in relation to the unit 'Education in the family'.

Sometimes a PTC heading as such had no equivalents in one or two of the target ILs but was provided with a 'contents description' (see Sect.1 of this article) containing some expressions which had their equivalents in the given target language (or languages). In this instance the equivalents of the said expressions were assigned to a given PTC heading. Let us illustrate this by the example of the following PTC heading:

0046 [record number]
04.19.00 Auxiliary sciences of history [PTC heading]
Contents description:
science of biography
genealogy
heraldry
sphragistics
numismatics
epigraphy
palaeography
diplomatics
historical criticism
TCT numbers [facet codes]: I.1.
Name of the TCT descriptor: SCIENCE OF BIOGRAPHY
Semantic equivalence to the PTC: >
UDC numbers: 929:736.3:737:930.27:930.22:930.1
UDC wordings:
SCIENCE OF BIOGRAPHY, GENEALOGY, HERALDRY
SPHRAGISTICS
NUMISMATICS
EPIGRAPHY, PALEOGRAPHY
DIPLOMATICS
HISTORICAL CRITICISM
Semantic equivalence to the PTC: =
Name of the SHL heading: AUXILIARY SCIENCES OF HISTORY
Semantic equivalence to the PTC: =

Various difficulties were encountered when assigning the lexical units of the target ILs to the PTC headings. The nature of these difficulties was to some extent different in the case of each of the target languages, but - in the main - they were caused by the insufficiency of the vocabularies of the target ILs and by imperfect or too restrictive grammar rules adopted in these languages.

Let us discuss in the first place some problems which arose when establishing the PTC -> TCT concordances. The PTC is - like most other classification systems - rather a pre-coordinated language while the TCT is characterized - just as other descriptor languages - by a rather high degree of post-coordination. So in many cases the contents of a very comprehensive PTC heading could be reflected only by a combination of two or more TCT descriptors (a grammar was adopted which consisted in forming the products of two or more descriptors by linking them by the sign ^; such a grammar was not used when carrying out the pilot studies). However, sometimes such combinations could not be formed because in the TCT some very general descriptors were lacking, such as theory, construction, methodology, etc. which could be easily combined with many other descriptors. E.g. the meaning of the PTC heading 45.09.00 'Theoretical problems of mining' could be expressed only by using the TCT descriptor MINING; the formation of the combination MINING^THEORY - which would denote almost exactly the meaning of the above-mentioned PTC heading - was not possible because the descriptor THEORY did not exist.

As far as the concordances between the PTC and the UDC are concerned, one of the hindrances was the lack of UDC numbers denoting some branches of activity. The names of these branches - appearing as PTC headings - could be expressed only inexactly by assigning the UDC numbers denoting some objects with which the given branches deal. E.g. to the PTC heading 68.35.21 'Hospital management' the UDC number 615.47 'Equipment of hospitals' was assigned, whereas the meaning of the PTC heading 66.31.00 'Hotel management' could be expressed only by assigning the UDC number 640.41 'Hotels. Pensions' (note: in the above examples the UDC wordings are given in an abbreviated form).

When establishing the concordances between the PTC and the third target language, i.e. the SHL, there was no possibility to denote - using the SHL language - such expressions as appear in the PTC tables as Prehistory, ancient history, Contemporary history (history of the twentieth century), General (universal) social and political history (social and political history of the world), General problems, present-day state, etc. This failure was caused by one of the specific features of the SHL - in this language only the use of standardized subject subheadings is allowed. Particularly painful here was the lack of certain subheadings having the form of adjectives (e.g. 'international') or expressing functions and processes (management, production, casting, melting, etc.). As a very illustrative example of the difficultiies encountered when establishing the PTC -> SHL concordances the problem of

expressing the concept of quality in the SHL language can be cited; in the list of subject-headings we can find the heading 'Quality' (which is narrower in relation to philosophy) as well as the headings 'Quality of production' and 'Quality of life' but there is no possibility to express the 'Quality of labour'. Another difficulty was caused by some of the scope notes inserted after subject subheadings; these notes narrowed the scopes of the subheadings after which they were inserted, e.g. after the subheading 'Programming' a note was placed which allowed the use of this subheading only after subject-headings relating to computer science.

All the above-mentioned difficulties had a negative influence on the results of the project, which are presented below.

3.2. Results of the project. Conclusions

The results obtained are shown in Table 8.

Target language (in relation to the PTC)	Full equivalence	Partial equivalence			Zero equivalence
	=	<	>	≠	0
TCT	402 30,23%	380 28,57%	74 5,56%	394 29,62%	80 6,02%
UDC	948 71,28%	83 6,24%	41 3,08%	228 17,14%	30 2,26%
SHL	699 52,56%	220 16,54%	38 2,86%	322 24,21%	51 3,83%

Table 8: Equivalence of TCT, UDC and SHL lexical units in relation to PTC headings.

In virtue of the results shown in Table 8 the occurrence of the equivalents of the PTC headings in the three target languages was investigated (see Table 9 below).

Kinds of equivalence	Occurence of the equivalence of the lexical units in the a, b, c, d languages						
	a b c d	a b c	a b d	a c d	a b	a c	a d
Full equivalence =	257	7	-	16	1	9	-
Partial equivalence <, >, ≠	949	24	18	37	3	1	2
Total	1206	31	18	53	4	10	2

a = PTC
b = TCT
c = UDC
d = SHL

Table 9: Compatibility of the four ILs under investigation (computed when establishing the CDIL)

On the basis of Table 9 the full and partial compatibility of the ILs under investigation was computed according to formula (1). The coincidence rate of all four languages a, b, c, d as well as the concidence rate between the master language a and each of the target languages b, c, d, were computed.

If only full equivalence was taken into consideration, the coincidence rate assumed the following values:

$$C_{abcd} = \frac{\sum_{i=1}^{1330} (a_i b_i c_i d_i)}{1330}$$

Thus:

$$C_{abcd} = \frac{257}{1330} = 0,193$$

$$C_{ab} = \frac{\sum_{i=1}^{1330} (a_i b_i)}{1330}$$

$$C_{ab} = \frac{265}{1330} = 0,199$$

$$C_{ac} = \frac{\sum_{i=1}^{1330} (a_i c_i)}{1330}$$

If full equivalence as well as partial equivalence was taken into account, the coincidence rate assumed the values as follows:

$$C'_{abcd} = \frac{1206}{1330} = 0,907$$

$$C'_{ab} = \frac{1259}{1330} = 0.947$$

$$C'_{ac} = \frac{1300}{1330} = 0,977$$

$$C'_{ad} = \frac{1279}{1330} = 0,962$$

Thus we see that if partial equivalence is taken into account and added to full equivalence, the coincidence rate increases in all cases almost five times. From the above results the following conclusions were drawn:

1) The highest degree of lexical compatibility occurs between the PTC and the UDC. From among the three target ILs the UDC is that indexing language in which appears the greatest number of lexical units which are the exact (full) equivalents of PTC headings (such exact equivalents could be assigned in more than 70% of cases); at the same time the UDC is characterized by the smallest number of failures, i.e. by the smallest number of cases of zero equivalence in relation to the PTC headings. Such a considerable degree of compatibility between the PTC and the UDC could be attained only by using in many cases the compound and complex UDC numbers and/or the multiple allocation of UDC numbers (i.e. by the assignment of two or more UDC numbers to one PTC heading). It seems that two main reasons of the rather high degree of compatibility between the PTC and the UDC can be pointed out, namely:

a) the exceptionally rich vocabulary of the UDC in comparison with the vocabularies of other ILs; b) the structural similarity between the PTC and the UDC - both ILs in question are hierarchical, discipline-oriented classification systems.

2) The SHL is characterized by a medium degree of compatibility with the PTC. In this language we could find a medium number of lexical units which were the exact equivalents of the PTC headings (more than 52% of the PTC headings had exact equivalents in the form of the SHL subject-heading entries). The SHL was also characterized by a medium number of cases of zero equivalence in relation to the PTC headings. Referring to the observations presented in chapter 3.1. of this article we can state that the degree of compatibility between the SHL and the PTC would be higher if some subject subheadings denoting activities, such as 'production', 'management', etc. were included in the vocabulary of the SHL.

3) The TCT is that target language which is characterized by the smallest number of cases (scarcely more than 30%) when the exact equivalents could be assigned to the PTC headings. At the same time the TCT is characterized by the greatest number of cases (more than 28%) when broader lexical units of a target language were assigned to the PTC headings. This is due to the fact that the TCT - unlike the other ILs under investigation (PTC, UDC and SHL) which are universal ones - has an interdisciplinary scope and many subject-fields - such as Philosophy, Mathematics, Physics, Chemistry, Biology, Medicine, Agriculture, Construction, Chemical industry, Food industry, etc. - which together occupy a considerable part of the PTC main table are represented in the TCT only to a very small extent, very often only by the name of the given subject-field and by the name of a few most important subfields. On the other hand the TCT contains many descriptors concerning some interdisciplinary problems; these descriptors have their equivalents in the PTC neither on the same level of hierarchy nor on the higher hierarchy levels. Generally speaking the extent to which the vocabularies of the PTC and the TCT overlap each other is not great. Referring to the considerations contained in Sect.3.1 of this article we can express the opinion that the degree of compatibility between the PTC and the TCT would considerable increase if a few very general descriptors, such as Theory, Construction, Methodology, etc. were added to the vocabulary of the TCT.

4) Further investigations concerning the comparison of the vocabularies of different ILs should be carried out. These investigations may e.g. concern the following problems: a) the possibility of expressing the scopes of the lexical units which are narrower in relation to the units under comparison when these units (having the same or similar meaning) belong to different hierarchical structures in the particular ILs under investigation; b) setting a certain degree of semantic distance betwen the lexical units under comparison (i.e. between a lexical unit of the master language and a unit belonging to the target language) beyond which we cannot admit the existence of any semantic relationship between these units (the compilers of the CDIL have noticed that the cases when zero equivalence occurs would be practically eliminated if very liberal criteria of semantic relationships were applied).

3.3 Applications of the CDIL already realized and possible future applications

The CDIL can first of all serve as a starting-point for the updating of the ILs which were taken into account when it was established. An attempt has already been made at using the Concordance Dictionary as a starting-point for the enlargement and refinement of the PTC. To this end the CDIL was first of all compared with the TCT. On the basis of this comparison a list of 1845 descriptors was established which had no equivalents in the form of PTC headings. Then this list of 1845 descriptors was confronted with the vocabularies of the PTC and the SHL. Thus we obtained a list of 380 expressions which appear in all three target languages but do not appear in the PTC tables. These expressions were admitted to be appropriate candidates for new PTC headings; records for these expressions were created in the PTC database.

In the opinion of its compilers the CDIL can also serve as:

1) a source of general and interdisciplinary vocabulary when elaborating branch ILs;
2) a tool wchich facilitates the indexing and re-indexing of documents by indexers-practitioners;
3) a didactic instrument permitting the demonstrative comparison of the ILs taken into account when the CDIL was elaborated;
4) a basis for future research (see also chapter 3.2. of this article).

4. Final remarks

The elaboration of the CDIL was a fascinating challenge to and a formidable experience for its compilers. They learned a lot about the nature of the ILs under investigation and of the ILs in general as well as about the possibilities and limitations connected with the establishment of concordances. Also the experience acquired in carrying out the pilot investigations was very useful.

A fragment of the original CDIL is shown in the Appendix to this article.

Appendix: Fragment of the Original CDIL (next page)

Notes

1 In paper (1) the previous version of this classification system was presented as the "Polish Subject-field Classification" (PSC).
2 All investigations described in this article were carried out on the basis of Polish terminology. For the sake of better understanding the examples were translated into English. However, the exact translation of Polish terms was not always possible.
3 The SHL does not use any notational system so in the case of this language only the verbal class description was assigned.

Appendix: Fragment of the Original CDIL

```
0620
41.00.00    INFORMATYKA
            Symbole TZW: I.1
            Nazwa deskryptora TZW: INFORMATYKA
            Odpowiednosć zakresowa z PKT: =
            Symbole UKD: 681.3
            Odpowiednik słowny UKD: INFORMATYKA
            Symbol odpowied. z PKT: =
            Nazwa SHP BN: INFORMATYKA
            Symbol odpowied. z PKT: =

0621
41.09.00    Podstawy teoretyczne informatyki
            Zawartosć hasła: - teoria komputerów
                             - projektowanie logiczne
                             - automatyzacja projektowania
            Symbole TZW: I.1., VI.4., VII.9.
            Nazwa deskryptora TZW: INFORMATYKA, KOMPUTER,
                                   AUTOMATYZACJA ^ PROJEKTOWANIE
            Odpowiednosć zakresowa z PKT: $
            Symbole UKD: 681.3.001
            Odpowiednik słowny UKD: TEORIA INFORMATYKI
            Symbol odpowied. z PKT: =
            Nazwa SHP BN: INFORMATYKA - TEORIA
            Symbol odpowiedn. z PKT: =

0622
41.15.00    Architektura komputerów
            Symbole TZW: VI.4
            Nazwa deskryptora TZW: KOMPUTER
            Odpowiednosć zakresowa z PKT: $
            Symbole UKD: 681.322.02
            Odpowiednik słowny UKD: ARCHITEKTURA KOMPUTERÓW
            Nazwa SHP BN: KOMPUTER
            Symbol odpowiedn. z PKT: $
```


References

(1) Scibor, E.: Polish subject-field classification - broad ordering system for use on a national scale. In: Perreault, J.M., Dahlberg, I. (Eds): Universal Classification Subject Analysis and Ordering Systems. Proc. 4th Int. Study Conf. on Classification Research, Augsburg, 28 June - 2 July 1982. Vol.2. Frankfurt: INDEKS Verlag 1983 p.133-8, 7 refs.

(2) Polska Klasyfikacja Tematyczna. III wersja. Oprac. Zespol Jezykow Informacyjnych Instytutu INTE pod kier. E. Scibora. Zaktualizowana edycja z 1991 r. Warszawa: IINTE 1991 (unpublished)

(3) Jabrzemska, E., Scibor, E: Zasady opracowywania i stosowania dziedzinowo-galeziowych klasyfikacji tematycznych. Warszawa: CINTE 1986. 54p = SINTO Materialy Metodyczne, No 26.

(4) Scibor, E., Jabrzemska, E.: Thesaurus of Common Topics. In: Kent, A: (Exec. Ed.): Encyclopedia of Libr.& Inform. Sci. Vol.44, Suppl.9. New York, NY: Marcel Dekker 1989, p.388-395, 7 refs.

(5) Bielicka, L.A., Paciejeswki, J., Scibor, E. (Comps.): Tezaurus Zagadnien Wspolnych - Wersja III. Czesc systematyczna (w ukladzie kategorialno-fasetowym). Czesc alfabetyczno-hierarchiczna. Warszawa: IINTE 1987 (unpublished).

(6) Uniwersalna Klasyfikacja Dziesietna. Wydanie skrocone dla polskiej sieci informacji naukowej. Konsorcjum ds UKD (UDCC). Publikacja nr UDC - POO5. T.1 \Tablice. Aktualne wg "Extensions and Corrections to the UDC" Ser. 14, No 3, 1992. Warszawa: IINTE 1993. 242 p.

(7) Trzcinska, J., Stepniakowa, E.: (Comps.): Slownik jezyka hasel przedmiotowych Biblioteki Narodowej. Stan na dzien 30 czerwca 1986r. Na podstawie "Slownika tematow dla bibliografij i katalogow w ukladzie przedmiotowym" J. Kossonogi oprac...Warszawa: Biblioteka Narodowa 1989. xxx,368 p.

(8) Trzcinska, J., Stepniakowa, E. (Comps.): Slownik jezyka hasel przedmiotowych Biblioteki Narodowej. Stan na dzien 31 grudnia 1992 r. Wyd. 2 por. i rozsz. T.1: A-L. XXV+382p. T.2: M-Z. 505p.. Warszawa: Biblioteka Narodowa 1993.

(9) Tomasik-Beck J., Analiza porownawcza wybranych jezykow informacyjnych o zakresie uniwersalnym lub interdyscyplinarnym. Prakt. i Teria Inform. Nauk. i Techn. 1994, No 3, p.11-16

(10) Dahlberg, I.: Toward establishment of compatibility between indexing languages. Int. Classif. 8(1981)No 2, p.86-91, 31 refs.

Prof.Dr. E. Scibor, Head of Department; Dr. J. Tomasik-Beck, Scientific Worker.

Department of Indexing Languages, Institute for Scientific, Technical and Economic Information, ul. Zurawia 4a, 00-503 Warszawa. POLAND

English translation of the Appendix in Polish given above:

APPENDIX
Fragment of the original CDIL (Translation into English)

[Record no] 0620
[PTC number] 41.00.00 COMPUTER SCIENCE
TCT numbers: I.1.
Name of the TCT descriptor: COMPUTER SCIENCE
Equivalence with PTC: =
UDC numbers: 681.3
Wording of the UDC: COMPUTER SCIENCE
Equivalence with PTC: =
Name of the SHL heading: COMPUTER SCIENCE
Equivalence with PTC: =

[Record no] 0621
[PTC number] 41.09.00 Theoretical foundations of computer science
Contents of the heading: - computer theory
- logical design
- design automation
TCT numbers: I.1., VI.4., VII.9
Name of the TCT descriptor: COMPUTER SCIENCE, COMPUTER, AUTOMATION^DESIGN
Equivalence with PTC: $
UDC numbers: 681.3.001
Wording of the UDC: THEORY OF COMPUTER SCIENCE
Equivalence with PTC: =
Name of the SHL heading: COMPUTER SCIENCE THEORY
Equivalence with PTC: =

[Record no] 0622
[PTC number] 41.15.00 Computer architecture
TCT numbers: VI.4.
Name of the TCT descriptor: COMPUTER
Equivalence with PTC: $
UDC numbers: 681.322.02
Wording of the UDC: COMPUTER ARCHITECTURE
Equivalence with PTC: =
Name of the SHL heading: COMPUTER
Equivalence with PTC: $