Godby, C., & Reighart, R. (1998). Using machine-readable text as a source of novel vocabulary to update the Dewy Decimal classification. 9th ASIS SIG/CR Classification Research Workshop, 91-106. doi:10.7152/acro.v9i1.12746

Using Machine-Readable Text as a Source of Novel Vocabulary to Update the Dewey Decimal Classification

Carol Jean Godby
Ray Reighart
OCLC Online Computer Library Center, Inc.

Abstract

A technique is presented for automatically importing novel and emergent vocabulary into the Dewey Decimal Classification (DDC) from a domain-specific corpus of machine-readable text using insights from recent work by computational linguistics on word sense disambiguation. Results show that most of the vocabulary is mapped to the DDC hierarchy that is most appropriate for characterizing the domain.

Introduction

Several research projects at OCLC are exploring the potential use of the Dewey Decimal Classification as a tool for classifying and browsing collections of digital information. As argued in Vizine-Goetz (1996), the DDC is well suited to this task because it is a general classification scheme with hierarchical structures and a rich notation for ordering classes. The DDC is already widely used in English-speaking countries and is well positioned to serve as a browsing and retrieval tool for accessing information across international borders. Translations that closely preserve the meaning of the Dewey notation as specified in the English language standard edition now exist for over 30 languages, including Spanish, Russian, Italian, and Mandarin Chinese.

Nevertheless, like any substantial reference work, the DDC must keep pace with change. The ongoing editorial processes required to update the DDC are discussed in Mitchell (1997, 1998). Experimental efforts to enhance the indexing vocabulary of the DDC by mapping Library of Congress Subject Headings are described in Vizine-Goetz (1997, 1998a). Our purpose here is to investigate the potential for importing additional indexing vocabulary to the DDC directly from machine-readable full text. This text documents the linguistic response to cultural change and is rapidly growing in volume and importance, as more full-text databases become available, as more electronic journals are published, and as the World Wide Web emerges as a new mass medium. Though the focus is on English-language text, the methods described here are theoretically extensible to other languages.

This project is a case study in a research program whose goal is to develop a library of generically useful software tools that identify subject-bearing terms and phrases and organize them by meaning, with or without reference to an external knowledge base such as the DDC (Godby 1998, Footnote 2). These tools should be able to facilitate access to stores of machine-readable text, primarily through better vocabulary identification, and can be used to create indexes and browse displays that more accurately represent an abstract view of the contents of a database. The same

ISSN: 2324-9773

Godby, C., & Reighart, R. (1998). Using machine-readable text as a source of novel vocabulary to update the Dewey Decimal classification. 9th ASIS SIG/CR Classification Research Workshop, 91-106. doi:10.7152/acro.v9i1.12746

tools developed to solve the practical problems of indexing a database can also be used to support projects in classification research because the starting point in such work is the identification of terminology that refers to new concepts or represents new names for concepts. This work as fundamentally interdisciplinary, drawing insights from linguistics, information retrieval, and information science.

Here we extend the work reported in Vizine-Goetz and Godby (1996) and Godby (1996), where we argued that current indexing vocabulary obtained from machine-readable text could be imported into the DDC using an information retrieval model. Standard corpus-linguistics techniques (Daille 1994) were used to extract novel phrases such as *virtual mall* or *Coldwell Banker* from contemporary Internet resources, along with a paragraph or two of local context. This context was further processed to obtain a vector of words and phrases that are highly associated with the phrase of interest. For the first target phrase, the highly associated words words and phrases from the local context included *Internet*, *online malls*, *marketplace*, *net*, *business*, and *shopping*; and for the second target, highly associated words and phrases included *agents*, *banker*, and *real estate*. The list of associates for each target phrase was submitted as a query to Scorpion (see Footnote 1), a Web-accessible automatic classification system developed at OCLC that uses relevant fields from the DDC as a SMART database (Salton 1971). The processes required for importing novel vocabulary to the DDC from unrestricted text are shown schematically in Figure 1.

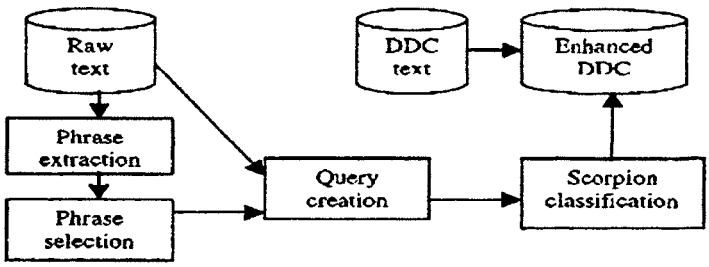

Figure 1. Process flow for mapping novel vocabulary obtained from free text to the DDC.

For a small set of hand-selected target phrases, the result was a list of DDC classes that suggested a reasonable domain of usage for the phrases. These DDC classes could, in turn, be enhanced with the target phrases as additional indexing vocabulary. Figure 2 shows the DDC classes returned by Scorpion for the two phrases.

ISSN: 2324-9773

Godby, C., & Reighart, R. (1998). Using machine-readable text as a source of novel vocabulary to update the Dewy Decimal classification. 9th ASIS SIG/CR Classification Research Workshop, 91-106. doi:10.7152/acro.v9i1.12746

Coldwell Banker		virtual mall	
346.04373	Closing and settlements	004.678	Internet
332.6324	Real estate	383.1	Retail trade
332.3322	Real estate market	380.1	Commerce (Trade)

Figure 2. Top three DDC classes returned by Scorpion for a query created by the local contexts of *Coldwell Banker* and *virtual mall*.

Figure 3 shows a portion of an experimental version of a DDC record created by Vizine-Goetz (1998b) for the class 306.8742 *Father-child relationship* and illustrates how the processes described in this paper for importing terminology into the DDC from unrestricted text complements other efforts to enhance the DDC. The phrase *fathers' rights* was added using the process described above. The Library of Congress Subject Headings are obtained by a procedure described in Vizine-Goetz (1998a) that extracts statistical associations between the DDC and the LCSH from bibliographic records contributed to OCLC's Online Union Catalog to which DDC numbers and LCSH indexing vocabulary have both been assigned.

Enhanced DDC records like those in Figure 3 can be used by the DDC editors to guide their decisions in writing captions or adding indexing vocabulary to future editions of the DDC. These records can also be converted to an electronic version of the DDC that will have more vocabulary to support the use of the DDC as an online browsing tool or to create more responsive Scorpion databases.

Class Number: 306.8742
Class Name: Father-child relationship

Dewey Hierarchy (structural information)

Broader Classes
300 Social sciences
306 Culture and institutions
306.8 Marriage and family
306.87 Intrafamily relationships
306.874 Parent-child relationship

Coordinate Classes
306.8742 Father-child relationship
306.8743 Mother-child relationship
306.8745 Grandparent-child relationship

Terminology

Dewey Index Terms
Father and child
Fatherhood
Fathers
Fathers--family relationships
Teenage fathers
Teenage fathers--family relationships
Unmarried fathers
Unmarried fathers--family relationships

Mapped Terms
LCSH/DDC (Dewey Web site)
Birthfathers
Free Text Terms
Fathers' rights
Statistically Associated LCSH
Fathers
Father and child
Fatherhood
Fathers and daughters

Figure 3. A DDC record with enhanced indexing.

ISSN: 2324-9773

Godby, C., & Reighart, R. (1998). Using machine-readable text as a source of novel vocabulary to update the Dewy Decimal classification. 9th ASIS SIG/CR Classification Research Workshop, 91-106. doi:10.7152/acro.v9i1.12746

Our current goal is to import novel vocabulary to the DDC on a large scale, using as a test collection a 60-megabyte corpus of recent news articles from the domain of politics and current affairs. We use the same general procedure described in our earlier work but have focused on making improvements to the selection of terms and the preprocessing of documents.

We are interested in vocabulary that is automatically selected, not hand selected, as in our previous work. Our goal is to identify phrases, usually compound nominals or noun phrases modified by adjectives, that have achieved word-like status in a particular subject. In a corpus of news stories about politics, these phrases include *alternative minimum tax*, *animal rights*, *ballistic missile defense*, *campaign finance reform* and *smart bombs* but not *old people*, *premature predictions*, *supply disruptions* and *welfare provisions*. We refer to the phrases of interest as `lexical phrases`.

Extracting lexical phrases from unrestricted text

When writers compose text, they must balance two countervailing forces. On the one hand, their knowledge of the rules of grammar permits them to be creative, putting words together to form phrases in ways that have never been done before and may only infrequently be done again. On the other hand, they have a vast knowledge of the conventional ways that the concepts in their chosen subject have been referred to before, including the use of lexical phrases such as *campaign finance reform*. Since our goal is to increase access to textual databases through improved searching and browsing tools, we are far more interested in conventional usage than a writer's creative use of language because conventional expressions include the vocabulary of a given subject that writers and searchers are likely to share. Subject-rich terminology that consists primarily of noun phrases has been attested by computational linguists working with many languages spoken in Europe, North America, and the Pacific Rim. For some recent studies, see Bourigault, Jacquemin and L'Homme (1998).

Computationally tractable methods for harvesting lexical phrases from unrestricted text must overcome the obstacle that their defining property cannot be practically formalized. Like words consisting of single terms, lexical phrases are words because they are names for persistent concepts. This property cannot be captured easily in a formal system because it encodes a relation between language and the world. For example, *school shooting* is a name for a persistent concept because the American press has recently been obsessed by a spate of gun-toting schoolboys who opened fire on their classmates. *Hotel shooting* has an analogous meaning but is far less common because hotels are less popular or remarkable venues for crime.

Nevertheless, lexical phrases, like single words, have many distributional properties that can be measured. First, relative to syntactic phrases, lexical phrases are more frequent and their component words are highly associated by standard measures such as mutual information (Church and Hanks 1990). For example, given that *missile* appears in a corpus of news articles, the likelihood is high that *ballistic* precedes it. The same point can be made about *animal rights*, *campaign finance*, and *Bill Clinton*, but not about *supply disruptions*, *premature predictions*, or *welfare provisions*.

ISSN: 2324-9773

Godby, C., & Reighart, R. (1998). Using machine-readable text as a source of novel vocabulary to update the Dewy Decimal classification. 9th ASIS SIG/CR Classification Research Workshop, 91-106. doi:10.7152/acro.v9i1.12746

Second, lexical phrases are highly specific to a single subject domain. Noun phrases such as *community development* or *direct government spending* are more commonly found in articles about political news than in articles about arts, health, or sports. However, phrases such as as *ample evidence*, *old people* or *vice president*, which might also be candidate lexical phrases because their component words are highly associated, appear in any number of contexts. This intuition was formalized in an algorithm by Zhou and Dapkus (1995), who isolated subject-rich phrases by comparing their relative frequencies in two corpora, one containing documents from a variety of subjects and one containing documents from a single subject.

Finally, lexical phrases form clusters based on syntactic similarity. Ranked frequency tabulations of noun phrase heads, the rightmost word in English phrases, successfully identify subject-rich lexical phrases in single-subject corpora and even in single documents, an observation that has been made for English (Godby 1998 and Wachtolder 1998), French (Ibekwe-San Juan 1998), and Japanese (Nakagawa and Kori 1998). In our experiments, the most common noun phrase heads in an English-language corpus about architecture were *architecture* and *design*, which appeared in phrases such as *landscape architecture* and *urban design*; in a corpus about metadata, the most common nominal heads were *element* and *type*, which appeared in the phrases *metadata element*, *author element*, *Internet media type*, and *resource type*.

Our procedure for identifying lexical phrases involves all three measures described above. The process flow is shown schematically in Figure 4.

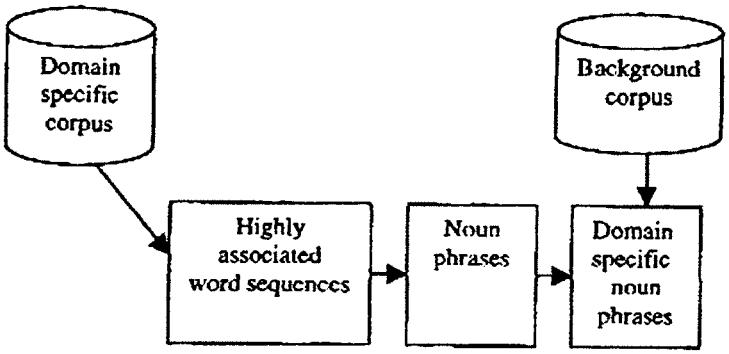

Figure 4. Process flow for obtaining domain-specific lexical phrases.

Using the processes outlined in Figure 1 on our test corpus, we collected 11878 lexical phrase candidates formed with 3987 different head nouns in an experimental run. The 15 most commonly occurring ones are listed in Table 1, along with counts of the lexical phrases containing them and some sample lexical phrases. Remarkably, just 15/3987, or approximately .04% of the head nouns, were the rightmost terms in over 14% of the lexical phrase candidates. This result supports the observation made by Nakagawa and Kori (1998) for a much smaller corpus of Japanese-language articles about computer science that many noun phrases in a domain-specific corpus share a relatively small number of unique heads.

ISSN: 2324-9773

Godby, C., & Reighart, R. (1998). Using machine-readable text as a source of novel vocabulary to update the Dewy Decimal classification. 9th ASIS SIG/CR Classification Research Workshop, 91-106. doi:10.7152/acro.v9i1.12746

Table 1. The fifteen most commonly occurring noun phrase heads in a corpus of political news.

Head	Lexical phrases containing head	Sample lexical phrase
program (-s)	238	affirmative action programs, corporate welfare programs, domestic spending program, housing programs, jobs programs
system (-s)	196	air defense system, ballistic missile system, child care system, criminal justice systems, democratic system, electoral system
policy (-ies)	143	agricultural policy, commercial policy, conservative policy, drug policy, energy policy, foreign policy, fiscal policy, global policy, technology policy
issue (-s)	125	abortion issue, campaign issues, character issue, foreign policy issues, Korean nuclear issue, prison labor issues
official (-s)	121	American trade officials, Forest Service officials, law enforcement officials, prison official
plan (-s)	119	affirmative action plan, benefit pension plan, Bosnian peace plan, employee pension plans, health plans, Medicare reform plan
committee (-s)	117	budget committee, campaign committee, domestic policy committee, House rules committee
reform (-s)	115	banking reform, campaign finance reform, health care reform, electoral reform, immigration reform, tort reform
party (-ies)	108	British Labour Party, Chinese Communist Party, Christian Democratic Pary, Communist Party, Grand Old Party, Liberal Democratic Party, tea party
government (s-)	107	American Government, big government, Blair Government, British Government, conservative government
tax (-es)	97	alternative minimum tax, aviation fuel tax, business tax, college tuition tax, corporate taxes, flat tax, wage tax, witholding tax
chairman (-men)	95	Federal Reserve Board Chairman, Democratic Party Chairman, Senate Banking Committee Chairman
act (-s)	87	American Dream Restoration Act, American Health Security Act, Balanced Budget Act, Budget Reconciliation Act, Clean Air Act, Emergency Food Assistance Act
leader (-s)	75	American leaders, black leaders, business leaders, evangelical leaders, legislative leaders, new GOP leaders
insurance	66	American Mutual Life Insurance, auto insurance, bank insurance, basic health insurance, casualty insurance, group health insurance, individual health insurance, medical insurance, no-fault auto insurance, private insurance, public health insurance, social insurance

The phrase selection processes described above allow at least three kinds of relatively low-risk mappings to the DDC.

• Mappings between real usage and controlled vocabulary. The lexical phrase *auto insurance* can be mapped to the DDC class 368.092 *Motor Vehicle Insurance. Automobile insurance*. Similarly, *health plan(s)* has been mapped to368.382 *Health Insurance,* and *flat tax* has been mapped to 336.205 *Tax Reform*, to which the LCSH *phrase flat-rate income tax--United States* has also been mapped by techniques described in Vizine-Goetz (1998a).

ISSN: 2324-9773

Godby, C., & Reighart, R. (1998). Using machine-readable text as a source of novel vocabulary to update the Dewy Decimal classification. 9th ASIS SIG/CR Classification Research Workshop, 91-106. doi:10.7152/acro.v9i1.12746

• Mappings of proper names to existing DDC categories. The DDC indexes already contain many proper names--not only the usual personal and geographic names, but also company names such as *Microsoft*. Names that are formed with commonly occurring head nouns, such as *American Mutual Life Insurance* and *Grand Old Party*, can be mapped to the DDC.

• Mappings of new concepts. The names for new concepts created by agents of cultural change are rarely coined words with no linguistic precedent. More commonly, they are lexical phrases that are formed by analogy with those already in common usage and whose component parts are probably represented in a modern classification system such as the DDC. Indeed, nearly all of the head nouns in Table 1 can be found in the DDC as captions or indexes that closely match the senses in which they are used in our test corpus. For example, the DDC class 368 *Insurance*, which contains index terms that enumerate types of insurance such as *mortgage insurance* and *prepaid health insurance*, could be augmented with the lexical phrases *no-fault auto insurance*, *basic health insurance*, and *individual health insurance*, all from our sample. As another example, the phrase *health care* exists in the DDC as a caption and an index entry, but *health care reform* and *health care reform legislation* do not, though *reform* and *legislation* are DDC classes that have been used by cataloging librarians to create custom DDC numbers for other topics of current interest such as *welfare reform*.

Importing lexical phrases to the DDC

To map lexical phrases to the DDC, we followed the same general procedure summarized in the Introduction and described more fully in Vizine-Goetz and Godby (1996) and Godby (1997). However, that procedure leaves considerable room for experimentation in the construction of queries. Intuitively, a query is simply a document containing the lexical phrase of interest that is submitted to the DDC version of the Scorpion database. If the query could be obtained from a canonical source, such as a dictionary definition or a paragraph in an encyclopedia, the DDC classes returned from Scorpion would be more likely to suggest reasonable domains of usage for the target phrases. But in our work, queries that simulate definitions of the target phrases must be constructed indirectly because the novel and emerging vocabulary that interests us is not defined in dictionaries and encyclopedias.

Our experiments with queries have been guided by recent results reported by computational linguists who have studied the problem of word sense disambiguation because the task of mapping new vocabulary to the DDC, however it is implemented, is an instance of the word sense disambiguation problem. As many have argued (see, for example, Yarowsky 1994), the word sense disambiguation problem has the same three essential parts that our problem has: words or lexical phrases whose senses or domains of usage are unknown, their citations or contexts of usage, and a lexical reference such as a dictionary, thesaurus, or classification system. A word is disambiguated when a similarity measure between the unknown word's contexts and a lexical reference results in a reasonable sense assignment. Consider, for example, the word *line*, a commonly studied word in the word sense disambiguation literature. A dictionary might list definitions for many senses, three of which are shown in Table 2:

ISSN: 2324-9773

Godby, C., & Reighart, R. (1998). Using machine-readable text as a source of novel vocabulary to update the Dewey Decimal classification. 9th ASIS SIG/CR Classification Research Workshop, 91-106. doi:10.7152/acro.v9i1.12746

Line 1 -- A theoretical mathematical construct that has length but no width.
Line 2 -- A fortified position, one marking the most forward position of troops.
Line 3 -- Electrical cable used to connect computers, telephones, or televisions to power stations.

Table 2. Three senses of the word *line*.

If all of the citations for *line* are drawn from appliance manuals, they are likely to contain words such as *electric*, *cable*, or *television* and a similarity measure that compares the citation to the definitions in Table 2 would choose *Line 3*.

For word sense disambiguation to succeed, not only must the citation selected for the target word characterize one of its senses, but the lexical reference must also provide disjoint definitions. In the dictionary fragment given in Table 2, the electrical, military, and geometric senses of *line* are distinguished clearly enough that even a simple similarity measure which counts matching words in a target word's citation and the sense definitions might be able to choose the correct result. The DDC's definitions are also disjoint, as demonstrated by Thompson, et al. (1997). When they used the DDC's own classes as queries against a DDC database, the top-ranked result was the query itself for 97% of the 31,000 queries.

The task of constructing queries for the lexical phrases in the public affairs corpus has been simplified by the vocabulary selection process described in the previous section.
Since only domain-specific phrases are selected, they are far less ambiguous than single-term words, such as *line,* that are usually studied by researchers interested in word sense disambiguation. As Yarowsky (1994) observed, "frozen" expressions such as the lexical phrases in Table 1 usually exhibit only one sense in a text, an observation that also extends to the component words in lexical phrases. In Table 1, the head nouns *program* and *system* have many sense definitions in standard dictionaries, but their usage in a domain-restricted corpus such as ours is remarkably consistent. Nevertheless, ambiguity still exists. Our 60-megabyte corpus contains approximately 7000 documents, in which concepts such as *health insurance* are discussed from economic, medical, political, and social policy perspectives.

One of the most robust results from recent studies of word sense disambiguation is that reliable cues for identifying a word's sense or domain usage can be obtained from the smallest local environments in a citation: the sentence or clause that contains the target word (see, for example, Yarowsky 1993, 1994). This can be seen trivially for personal names in news articles, where titles and appositives are prevalent, as the examples from our data in Table 3 show:

Arizona senator John McCain
Attorney general Janet Reno
Deputy budget director Alice Rivlin
broadcaster Pat Robinson
cranky liberal Barney Frank
Syrian president Hafez Assad
White House spokesman Mike McCurry
Wisconsin governor Tommy Thompson

Table 3. Local contexts for personal names from a corpus of political news.

ISSN: 2324-9773

Godby, C., & Reighart, R. (1998). Using machine-readable text as a source of novel vocabulary to update the Dewy Decimal classification. 9th ASIS SIG/CR Classification Research Workshop, 91-106. doi:10.7152/acro.v9i1.12746

Local contexts for lexical phrases whose head nouns are listed in Table 1 also show that the immediate syntactic context contains many clues that identify relevant perspectives. In sentences 1 and 2 in Table 4, the co-occurrence of *tax code, tax break* and *economies of scale* suggest that *health insurance* is discussed from an economic perspective. In sentences 3 and 4, the co-occurence of *government mistakes, state-run,* and *legislature* are clues that *health insurance* is discussed from a political perspective.

1....those views include changing the tax code so that individuals get the same tax breaks as employers when they buy health insurance...
2...the largest employers would purchase their health insurance, providing economies of scale...
3...most of the problems of the current system are attributable to prior government mistakes that have led us to purchase too much of the wrong kind of health insurance... .
4...significant majorities of the public favor reinforced policing of crime, national health insurance, and greater economic security
their attention soon fixed on the legislature, which a few years before had ignored repeated warnings of flaws in the state-run insurance system...

Table 4. Local contexts for *health insurance* in a corpus of political news

Following the recommendations in Grefenstette (1994), we collected all citations in the corpus containing the local syntactic context for each lexical phrase that to be added to the DDC. The contexts were parsed to extract noun phrases and verb phrases and submitted as queries to Scorpion.

Some lexical phrases in our corpus, such as *health insurance*, are already present in the DDC as captions or indexing vocabulary. These are valuable test cases because the classes to which they have been assigned by the DDC's editors can serve as convenient relevance judgments for information retrieval experiments on our system. To test the value of the additional processing described in this section, we constructed two sets of queries for a set of 50 such phrases. In the baseline condition, a query consisted of a single paragraph of raw text surrounding the target phrase. In the test condition, the citation was a single sentence surrounding the target phrase from which noun and verb phrases were extracted. For both conditions, recall was near 90% for 20-item result sets. But in the test condition, approximately 60% of the relevant classes were among the top five results; in the baseline condition, only 20% were in the top five. These results suggest that the effect of more sophisticated preprocessing of queries is to move the DDC class assignments selected by human editors for the target phrases to the top of the result set, thus simplifying the challenging problem of filtering correct answers from inevitable noise.

Using the most successful procedure for constructing queries, we also experimented with the mapping of novel lexical phrases. Though relevance judgments do not exist for the novel lexical phrases, it is possible to gain insight into the performance of our system by examining the overall distribution of the Scorpion assignments. In preliminary tests, we automatically extracted and created queries for 272 of the most frequent phrases created from frequently occurring head nouns. As in previous tests, the queries were submitted to Scorpion to obtain DDC class

ISSN: 2324-9773

Godby, C., & Reighart, R. (1998). Using machine-readable text as a source of novel vocabulary to update the Dewy Decimal classification. 9th ASIS SIG/CR Classification Research Workshop, 91-106. doi:10.7152/acro.v9i1.12746

assignments. Table 5 shows the distribution of the DDC-100 classes for the top-ranked result for each query. Nearly 88% of the assignments were to *300 Social sciences*, the DDC hierarchy that would best characterize a corpus of political news.

DDC Number	DDC Caption	Count	Percent
000	Generalities	2	.007
100	Philosophy, paranormal phenomena, psychology	3	.011
200	Religion	8	.029
300	Social sciences	239	.878
400	Langugage	0	—
500	Natural sciences and mathematics	1	.003
600	Technology (applied sciences)	15	.055
700	The arts	2	.007
800	Literature (Belles-lettres) and rhetoric	2	.007
900	Geography, history, and auxiliary disciplines	0	—

Table 5. The DDC-100 class assignments of the top-ranked results for queries that characterize 272 lexical phrases obtained from a corpus of news articles about political affairs

Table 6 shows the distribution of the 239 top-ranked results that were assigned to the *Social Sciences* hierarchy. Over 80% of the novel lexical phrases were assigned to just four sub-hierarchies, *320 Political science*, *330 Economics*, *340 Law*, and *350 Public administration and military science.*

DDC Number	DDC Caption	Count	Percent
300	Social science	6	.025
310	Collections of general statistics	0	–
320	Political science	74	.309
330	Economics	51	.213
340	Law	36	.150
350	Public administration and military science	31	.129
360	Social problems and services	19	.079
370	Education	11	.046
380	Commerce, communications, transportation	11	.046
390	Customs, etiquette, folklore	0	–

Table 6. Distribution of 239 top-ranked results assigned to the DDC 300 hierarchy for queries that characterize 272 lexical phrases obtained from a corpus of news articles about political affairs

Tables 7 and 8 show the same tight distribution of DDC class assignments in the top ten results returned for each query. Even with a result set that is an order of magnitude larger, nearly 84% of the assignments are still in the DDC hierarchy *300 Social sciences*.

ISSN: 2324-9773

Godby, C., & Reighart, R. (1998). Using machine-readable text as a source of novel vocabulary to update the Dewy Decimal classification. 9th ASIS SIG/CR Classification Research Workshop, 91-106. doi:10.7152/acro.v9i1.12746

DDC Number	DDC Caption	Count	Percent
000	Generalities	43	.016
100	Philosophy, paranormal phenomena, psychology	35	.013
200	Religion	111	.040
300	Social sciences	2278	.837
400	Langugage	4	.001
500	Natural sciences and mathematics	23	.008
600	Technology (applied sciences)	186	.068
700	The arts	27	.010
800	Literature (Belles-lettres) and rhetoric	13	.005
900	Geography, history, and auxiliary disciplines	0	—
Total		2720	

Table 7. The DDC-100 class assignments of the top ten results for queries that characterize 272 lexical phrases obtained from a corpus of news articles about political affairs

Table 8 shows that even when larger result sets are examined, the lexical phrases are still primarily assigned to four of the ten subclasses in the Social Sciences hierarchy.

DDC Number	DDC Caption	Count	Percent
300	Social science	47	.020
310	Collections of general statistics	0	–
320	Political science	654	.287
330	Economics	338	.148
340	Law	434	.190
350	Public administration and military science	382	.167
360	Social problems and services	201	.088
370	Education	137	.060
380	Commerce, communications, transportation	80	.035
390	Customs, etiquette, folklore	5	.002
Total		2278	

Table 8. Distribution of 2278 top ten results assigned to the DDC 300 hierarchy for queries that characterize 272 lexical phrases obtained from a corpus of news articles about political affairs

Taken together, the data in Tables 5-8 provide evidence that DDC's class definitions are disjunct and support the argument that local syntactic context works remarkably well in identifying powerful clues for the domain usage of novel lexical phrases. Of course, the real test is the quality of the individual mappings, which can be evaluated only by classification experts or users who find the DDC enhanced with the lexical phrases introduced by the processes described in this paper valuable. Table 9 shows the top-ranked class returned by Scorpion for 20 lexical phrases.

ISSN: 2324-9773

Godby, C., & Reighart, R. (1998). Using machine-readable text as a source of novel vocabulary to update the Dewy Decimal classification. 9th ASIS SIG/CR Classification Research Workshop, 91-106. doi:10.7152/acro.v9i1.12746

DDC Class	DDC Caption	Upward Hierarchy		Lexical Phrase
305.3	Men and women	305	Social groups	gender gap, gender preferences
320.5322	Marxism-Leninism	320.53	Collectivism and fascism	Evil Empire
324.274704	Republican party	324.2	Political parties	Grand Old Party
333.72	Conservation and protection	333.7	Natural resources and energy	Clean Air Act
336.243	Corporate income taxes	336.24	Income taxes	corporate welfare, corporate welfare program, alternative minimum tax
336.39	Expenditure	336.3	Public borrowing, debt, expenditure	jobs programs, domestic sspending programs
342.03	Revision and amendment of the basic instruments of government	342	Constitutional and administrative law	human life amendment
353.27	Air defense system	353.2	Military and defense administration	air defense system
353.54	Income maintenance	353.5	Administration of social welfare	social insurance
353.690973	Health insurance	353.6	Administration of health services	Medicare reform plan
361.24	Reform movements	361.2	Social action	immigration reform
363.58	Programs and services	363.5	Housing	housing programs
368.382	Health insurance	368	Insurance	health plan
368.4200973	Accident and health insurance (Medicaid)	368	Insurance	basic health insurance
658.3253	Pensions	358.32	Wage and salary administration	employee pension plans
801.9	Nature and character	801	Philosophy and theory	character issue

Table 9 shows the top-ranked class returned by Scorpion for 20 lexical phrases

Applications

With the criteria we have defined to select phrases and create queries, it is feasible to map large amounts of vocabulary that represents current English usage to the DDC. This data can be used in a variety of ways. Our original intent was to add end-user indexing vocabulary to the DDC, supporting and extending the work of Vizine-Goetz (1996).

But the DDC can also be used in the background as an implicit structure that filters and supplements the automatic thesauri that can be generated from scratch by the information-retrieval and corpus-linguistics techniques explored by researchers such as Grefenstette (1994) and Croft and Yufeng (1994). Finally, our data is independently useful to the editors of the DDC because it automates some of the work performed by lexicographers.

ISSN: 2324-9773

Godby, C., & Reighart, R. (1998). Using machine-readable text as a source of novel vocabulary to update the Dewy Decimal classification. 9th ASIS SIG/CR Classification Research Workshop, 91-106. doi:10.7152/acro.v9i1.12746

Acknowledgements

We wish to thank the anonymous reviewer, whose comments resulted in many improvements to the style and substance of this paper.

Footnotes

1. Scorpion is accessible at: http://orc.rsch.oclc.org:6109/
2. The WordSmith project is accessible at: http://orc/WordSmith/wshome.html

References

Bourigault, D., Christian Jacquemin C., and L'Homme, M. (1998, August). *Computerm '98: Proceedings from the First Workshop on Computational Terminology.* Montreal, Quebec, Canada: COLING-ACL.

Church, K., and Hanks, P. (1990). Word association norms, mutual information,and lexicography. *Computational Linguistics 16:1*, 22-29.

Croft, B., and Yufeng, J. (1994). An association thesaurus for information retrieval. *Intelligent Multimedia Information Retrieval Systems & Management: Proceedings of the RIAO Conference.* New York, New York. 146-160.

Daille, B. (1994). Study and implementation of combined techniques for automatic extraction of terminology. *Proceedings from the Balancing Act.* Las Cruces, New Mexico: the Association for Computational Linguistics. 29-36.

Godby, C. (1996). Enhancing the indexing vocabulary of the Dewey Decimal Classification. In the *Annual Review of OCLC Research.*
Accessible at http://www.oclc.org/oclc/research/publications/review96/frames_man.htm

Godby, C. (1998a). The WordSmith Toolkit. The *Annual Review of OCLC Research.* Accessible at: http://www.oclc.org/oclc/research/publications/review97.

Godby, C. (1998b). English compound nominals in context. Unpublished manuscript. The Ohio State University.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery.* Boston: Kluwer Academic Press.

Ibekwe-San Juan, F. (1998). Building a prototype system for trends survey through the use of syntactic term variants. *Computerm '98: Proceedings from the First Workshop on Computational Terminology.* Montreal, Quebec, Canada: COLING-ACL. 22-28.

ISSN: 2324-9773

Godby, C., & Reighart, R. (1998). Using machine-readable text as a source of novel vocabulary to update the Dewey Decimal classification. 9th ASIS SIG/CR Classification Research Workshop, 91-106. doi:10.7152/acro.v9i1.12746

Johnston, M., Boguraev, B., and J. Pustejovsky, J. (1995). The acquisition and interpretation of complex nominals. *Working notes of AAAI95 Spring Symposium on the Representation and Acquisition of Lexical Knowledge.* Accessible at: http://www.cs.brandeis.edu/~rllc/pub.html

Mitchell, J. (1997). DDC 21: An Introduction. In *Dewey Decimal Classification Edition 21 and International Perspectives*. Albany, New York: Forest Press, 3-15.

Mitchell, J. (1998, March/April). Dewey used around the world. *OCLC Newsletter,* 25-27.

Nakagawa, H. and Mori, T., (1998). Nested collocation and compound noun for term extraction.). *Computerm '98: Proceedings from the First Workshop on Computational Terminology.* Montreal, Quebec, Canada: COLING-ACL. 64-70.

Salton, G.(Ed.) (1971). *The SMART Retrieval System -- Experiments in Automatic* Document Processing. Eglewood Cliffs, N.J.: Prentice Hall, Inc.

Thompson, R., Shafer, K., and Vizine-Goetz, D. (1997). Evaluating Dewey concepts as a knowledge base for automatic subject assignment. Presented at the First Digital Library Conference, Philadelphia. Accessible at: http://orc.rsch.oclc.org:6109/
http://orc.rsch.oclc.org:6109/eval_dc.html.

Vizine-Goetz., D. (1996). Classification research at OCLC. In the *Annual Review of OCLC Research*. Accessible at: http://www.oclc.org/oclc/research/publications/review96/frames_man.htm

Vizine-Goetz, D. (1997). Popular LCSH with Dewey numbers (subject headings for anyone). In *Annual Review of OCLC Research 1997.* Accessible at: htttp://www.oclc.org/oclc/research/publications/review97.

Vizine-Goetz, D. (1998a May/June). Subject headings for everyone: popular Library of Congress subject headings with Dewey numbers. *OCLC Newsletter,* 29-33.

Vizine-Goetz, D. (1998b). Dewey as an Internet subject guide. Manuscript in preparation.

Vizine-Goetz. D. and Godby, C. (1996). Library classification schemes and access to electronic collections: enhancement of the Dewey Decimal Classification with supplemental vocabulary. *Advances in Classification Research Volume 7: Proceedings of the 7th ASIS SIG/CR Classification Research Workshop,* Silver Spring, MD: American Society for Information Science.

Wacholder, N. (1998). Simplex NPs clustered by head: a method for identifying significant topics within a document. *The Computational Treatment of Nominals: Proceedings of the Workshop*. Montreal, Quebec, Canada: COLING-ACL. 70-79.

ISSN: 2324-9773

Godby, C., & Reighart, R. (1998). Using machine-readable text as a source of novel vocabulary to update the Dewy Decimal classification. 9th ASIS SIG/CR Classification Research Workshop, 91-106. doi:10.7152/acro.v9i1.12746

Yarowsky, D. (1993). One sense per collocation. *In Proceedings of the ARPA Workshop on Human Language Technology*. San Francisco: ARPA Software and Intelligent Systems Technology Office, Morgan Kaufmann, 266-71.

Yarowsky, D. (1994). Decision lists for lexical ambiguity resolution: application to accent restoration in Spanish and French. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, 88-95.

Zhou, J. and Dapkus, P. (1995). Automatic suggestion of significant terms for a predefined topic. In *Proceedings of the Third Workshop on Very Large Corpora*. 131-147.

ISSN: 2324-9773

Godby, C., & Reighart, R. (1998). Using machine-readable text as a source of novel vocabulary to update the Dewy Decimal classification. 9th ASIS SIG/CR Classification Research Workshop, 91-106. doi:10.7152/acro.v9i1.12746

ISSN: 2324-9773