Revising the WORDNET DOMAINS Hierarchy: semantics, coverage and balancing

Luisa Bentivogli, Pamela Forner, Bernardo Magnini, Emanuele Pianta
ITC-irst – Istituto per la Ricerca Scientifica e Tecnologica
Via Sommarive 18, Povo – Trento, Italy, 38050
email:{bentivo, forner, magnini, pianta}@itc.it

Abstract

The continuous expansion of the multilingual information society has led in recent years to a pressing demand for multilingual linguistic resources suitable to be used for different applications.

In this paper we present the WordNet Domains Hierarchy (WDH), a language-independent resource composed of 164, hierarchically organized, domain labels (e.g. Architecture, Sport, Medicine). Although WDH has been successfully applied to various Natural Language Processing tasks, the first available version presented some problems, mostly related to the lack of a clear semantics of the domain labels. Other correlated issues were the coverage and the balancing of the domains. We illustrate a new version of WDH addressing these problems by an explicit and systematic reference to the Dewey Decimal Classification. The new version of WDH has a better defined semantics and is applicable to a wider range of tasks.

1 Introduction

The continuous expansion of the multilingual information society with a growing number of new languages present on the Web has led in recent years to a pressing demand for multilingual applications. To support such applications, multilingual language resources are needed, which however require a lot of human effort to be built. For this reason, the development of language-independent resources which factorize what is common to many languages, and are possibly linked to the language-specific resources, could bring great advantages to the development of the multilingual information society.

A language-independent resource, usable in many automatic and human applications, is represented by *domain hierarchies.* The notion of domain is related to similar notions such as *semantic field*, *subject matter*, *broad topic*, *subject code*, *subject domain*, *category*. These notions are used, sometimes interchangeably, sometimes with significant distinctions, in various fields such as linguistics, lexicography, cataloguing, text categorization. As far as this work is concerned, we define a *domain* as an area of knowledge which is somehow recognized as unitary. A domain can be characterized by the name of a discipline where a certain knowledge area is developed (e.g. chemistry) or by the specific object of the knowledge area (e.g. food). Although objects of knowledge and disciplines that study them are clearly related, the relation between these two points of view on domains is sometimes blurred and may be a source of uncertainty on their exact definition.

Another interesting duality when speaking about domains is related to the fact that knowledge manifests itself in both words and texts. So the notion of domain can be applied both to the study of words, where a domain is the area of knowledge to which a certain lexical concept belongs, or to the study of texts, where the domain of a text is its broad topic. In this work we will assume that also these two points of view on domains are strictly intertwined.

By their nature, domains can be organized in hierarchies based on a relation of specificity. For instance we can say that TENNIS is a more specific domain than SPORT, or that ARCHITECTURE is more general than TOWN PLANNING.

Domain hierarchies can be usefully integrated into other linguistic resources and are also profitably used in many Natural Language Processing (NLP) tasks such as Word Sense Disambiguation (Magnini et al. 2002), Text Categorization (Schutze, 1998), Information Retrieval (Walker and Amsler, 1986).

As regards the usage of Domain hierarchies in the field of multilingual lexicography, an example is given by the EuroWordNet Domain-ontology, a language independent domain hierarchy to which interlingual concepts (ILI-records) can be assigned (Vossen, 1998). In the same line, see also the SIMPLE domain hierarchy (SIMPLE, 2000).

Large domain hierarchies are also available on the Internet, mainly meant for classifying web documents. See for instance the Google and Yahoo directories.

A large-scale application of a domain hierarchy to a lexicon is represented by WORDNET DOMAINS (Magnini and Cavaglià, 2000). WORDNET DOMAINS is a lexical resource developed at ITC-irst where each WordNet synset (Fellbaum, 1998) is annotated with one or more domain labels

selected from a domain hierarchy which was specifically created to this purpose. As the WORDNET DOMAINS Hierarchy (WDH) is language-independent, it has been possible to exploit it in the framework of MultiWordNet (Pianta et al., 2002), a multilingual lexical database developed at ITC-irst in which the Italian component is strictly aligned with the English WordNet. In MultiWordNet, the domain information has been automatically transferred from English to Italian, resulting in a Italian version of WORDNET DOMAINS. For instance, as the English synset {court, tribunal, judicature} was annotated with the domain LAW, also the Italian synset {corte, tribunale}, which is aligned with the corresponding English synset, results automatically annotated with the LAW domain. This procedure can be applied to any other WordNet (or part of it) aligned with Princeton WordNet (see for instance the Spanish WordNet).

It is worth noticing that two of the main on-going projects addressing the construction of multilingual resources, that is MEANING (Rigau et al. 2002) and BALKANET (see web site), make use of WORDNET DOMAINS. Finally, WORDNET DOMAINS is being profitably used by the NLP community mainly for Word Sense Disambiguation tasks in various languages.

Another application of domain hierarchies can be found in the field of *corpus creation*. In many existing corpora (see for instance the BNC, the ANC, the Brown and LOB Corpora) domain is one of the most used criteria for text selection and/or classification. Given that a domain hierarchy is language independent, if the same domain hierarchy is used to build reference corpora for different languages, then it would be easy to create (a first approximation of) *comparable corpora* by putting in correspondence corpora sections belonging to the same domain.

An example of a corpus in which the complete representation of domains is pursued in a systematic way is represented by the MEANING Italian corpus, a large size corpus of written contemporary Italian in which a subset of the WDH labels has been chosen as the fundamental criterion for the selection of the texts to be included in the corpus (Bentivogli et al., 2003).

Given the relevance of language-independent domain hierarchies for multilingual applications, it is of primary importance that these resources have a well-defined semantics and structure in order to be useful in various application fields. This paper reports the work done to improve the WDH so that it complies with such requirements. In particular, the WDH revision has been carried out with reference to the Dewey Decimal Classification.

The paper is organized as follows. Section 2 briefly introduces the WORDNET DOMAINS Hierarchy and its main characteristics, with a short overview of the Dewey Decimal Classification system. Section 3 describes features and properties of the revision. Finally, in section 4, conclusions are reported.

2 The WordNet Domains Hierarchy

The first version of the WDH was composed of 164 domain labels selected starting from the subject field codes used in current dictionaries, and the subject codes contained in the Dewey Decimal Classification (DDC), a general knowledge organization tool which is the most widely used taxonomy for library organization purposes.

Domain labels were organized in five main trees, reaching a maximum depth of four. Figure 1 shows a fragment of one of the five main trees in the WORDNET DOMAINS original hierarchy.

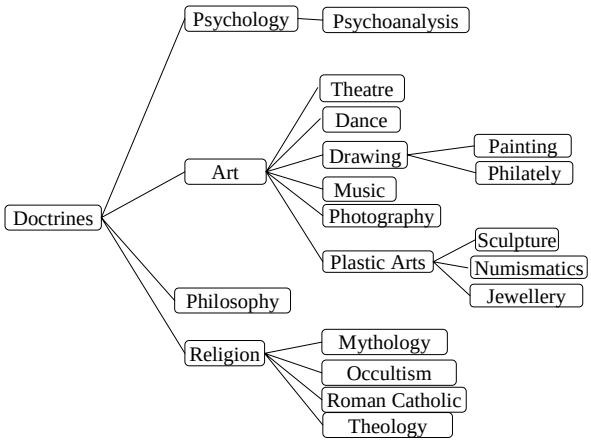

Figure 1: Fragment of the original WDH

Domain labels were initially conceived to be application-oriented, that is, they have been integrated in WordNet with the main purpose of allowing the categorization of word senses and to provide useful information during the disambiguation process.

The second level of WDH, where the so-called *Basic Domains* are represented, includes labels such as ART, SPORT, RELIGION and HISTORY, while in the third level a degree of major specialization is reproduced, and domains, like for example, DRAWING, PAINTING, TENNIS, VOLLEYBALL, and ARCHAEOLOGY can be found. For NLP tasks, the set of *Basic Domains* has proved to possess a suitable level of abstraction and granularity.

Although the first version of WDH found many applications in different scenarios, it presented some problems. First, the domain labels did not have a defined semantics. The content of the labels

could be suggested by the lexical meaning of their name, but there was no explicit indication about their intended interpretation.

Second, it was not clear whether the *Basic Domains* met certain requirements such as knowledge coverage and balancing. In fact, the *Basic Domains* are supposed to possess a comparable degree of granularity and, at the same time, to cover all human knowledge. However, they did not always posses such characteristics. For instance VETERINARY was put at the same level as ECONOMY, although these two domains obviously do not posses the same level of granularity. Moreover not all branches of human knowledge were represented (see for instance the HOME domain).

The purpose of the work presented here was, therefore, to find a solution for such problems, in order to improve the applicability of WDH in a wider range of fields. The solution we propose is crucially based on the Dewey Decimal Classification (edition 21), which has been used as a reference point for defining a clear semantics, preventing overlapping among domains, and assessing the *Basic Domains* coverage and granularity issues.

2.1 The Dewey Decimal Classification (DDC)

The Dewey Decimal Classification (DDC) system (Mitchell et al. 1996) is the most widely used taxonomy for library classification purposes providing a logical system for the organization of every item of knowledge through well-defined subject codes hierarchically organized. The semantics of each subject code is determined by a numeric code, a short lexical description associated to it, and by the hierarchical relations with the other subject codes. Another characteristic of the DDC is that a handbook is available explaining how texts should be classified under subject codes.

The DDC is not just for organizing book collections; it has also been licensed for cataloguing internet resources (see for example BUBL http://bubl.ac.uk/link/) and it was conceived to accommodate the expansion and evolution of the body of human knowledge.

The DDC hierarchy is arranged by disciplines (or fields of study), and this entails that a subject may appear in more than one discipline, depending on the aspect of the topic discussed.

The DDC hierarchical structure allows a topic to be defined as part of the broader topic above it, and that determines the meaning of the class and its relation to other classes. At the broadest level, called *Main Classes* (or *First summary)*, the DDC is composed of ten mutually exclusive main classes, which together cover the entire world of knowledge. Each main class is sub-divided into ten *divisions*, (the *Hundred Divisions*, or *Second Summary*) and each division is split into ten *sections* (the *Thousand Section*, also called *Third Summary*).

Each category in the DDC is represented by a numeric code as the example below shows.

700 *Art*
- 7**3**0 *Plastic Arts*
  - *73**6** Carving*
    - *736.**2** Precious Stones*
      - *736.2**3** Diamonds*
      - *736.2**5** Sapphires*
    - *736.**4** Wood*
  - *73**8** Ceramic Arts*
  - ***73**9** Art Metalwork*
- 7**4**0 *Drawing*
- 7**5**0 *Painting*

The first digit of the numbers indicates the main class, (**7**00 is used for all *Arts*) the second digit indicates the hundred division, (7**3**0 corresponds to *Plastic arts*, 7**4**0 to *Drawing*, 7**5**0 to *Painting*) and the third digit indicates the section (73**6** represents *Carving*, 73**8** *Ceramic arts*, 73**9** *Art metalwork*). Moreover, almost all sub-classes are further subdivided. A decimal point follows the third digit until the degree of specification needed (736.2**3** *Diamonds*, 736.2**5** *Sapphires*).

3 The Revision of the WDH

The revision of the first version of the WDH aimed at satisfying the following properties and characteristics:

- *semantics*: each WDH label should have an explicit semantics and should be unambiguously identified;
- *disjunction*: the interpretation of all WDH labels should not overlap;
- *basic coverage*: all human knowledge should be covered by the *Basic Domains*;
- *basic balancing*: most *Basic Domains* should have a comparable degree of granularity.

In the following sections we are going to show how a systematic mapping between WDH and DDC can be used to enforce each of the above characteristics.

3.1 Semantics

To give the domain labels a clear semantics so that they can be unambiguously identified and interpreted, we decided to associate each domain label to one or more DDC codes as shown below in Table 1.

WDH Domains	DDC Codes
Art	[700-(790-(791.43,792,793.3), 710,720,745.5)]
Plastic arts	730
Sculpture	[731:735]
Numismatics	737
Jewellery	739.27
Drawing	[740-745.5]
Painting	750
Graphic arts	760
Philately	769.56
Photography	770
Music	780
Cinema	791.43
Theatre	[792-792.8]
Dance	[792.8,793.3]

Table 1: Fragment of the new WDH with the respective DDC codes

In many cases we found a one-to-one mapping between a WDH label and a DDC code (e.g. PAINTING mapped onto 750 or CINEMA onto 791.43). When one-to-one mappings were not found, artificial DDC codes were created. An artificial code, represented within square brackets, is created with reference to various DDC codes or parts of them. To describe artificial nodes, certain conventions have been adopted.

(i) A series of non-consecutive codes is listed separated by a comma (see DANCE).

(ii) A series of consecutive codes is indicated by a range. For instance, the series [731, 732, 733, 734, 735] is abbreviated as [731:735] (see SCULPTURE).

(iii) A part of a tree is represented as the difference between a tree and one or more of its subtrees, where the tree and the subtrees are identified by their roots (see DRAWING).

(iv) The square brackets should be interpreted as meaning "the generalities" of the composition of codes contained in the brackets. So, for instance, [731:735] should be interpreted as the generalities of the codes going from 731 to 735. In the original DDC, generalities are identified by the 0 decimal. For instance, the code 700 refers to the generalities of the codes from 710 to 790.

To establish a mapping between labels and codes we exploited the names of the DDC categories and their description in the DDC manual. This worked pretty well in most cases, but there are some exceptions. Take for instance the TOURISM domain. Apparently tourism does not occur as a category in the DDC. On a closer inspection it came out that the categories which are most clearly related to tourism are 910.202:*World travel guides* and 910.4:*Accounts of travel*.

Note that a WDH domain can be mapped onto codes included in different DDC main classes, i.e. disciplines. For example ARTISANSHIP (745.5:*Handicrafts*, 338.642:*Small business*) maps onto categories located partly under 700:*Art* and partly under 300:*Social Sciences*. The same happens with SEXUALITY, a domain that following the DDC is studied by many different disciplines, e.g. philosophy, medicine, psychology, body care.

As a consequence of the systematic specification of the semantics of the WDH domains, some of them have been re-labeled with regard to the previous version of the hierarchy. For instance, the domain BOTANY has been changed to PLANTS, ZOOLOGY to ANIMALS, and ALIMENTATION to FOOD. This change of focus from the name of the discipline to the name of the object of the discipline is not only in compliance with the new edition of the DDC, but it also reflects current and international usage (see, for example, Google categories). In some cases the change of the domain name comes along with a change of its intended interpretation. For instance, we have decided to enlarge the semantics of the domain ZOOTECHNICS and to call it ANIMAL HUSBANDRY, a more generic domain which was missing in the previous hierarchy.

In most cases the hierarchical relations between the WDH domains are the same as the relations holding between the corresponding DDC codes: MUSIC is more specific than ART in the same way as 780:*Music* is more specific than 700:*The Arts*. To reinforce the hierarchical parallelism between the WDH and the DCC, we re-located some domains with regard to the previous WDH hierarchy. For example, OCCULTISM, which was placed under RELIGION in the old hierarchy, has been moved under the newly created domain PARANORMAL. Also, TOPOGRAPHY, previously placed under ASTRONOMY, has now been moved under GEOGRAPHY.

In a few cases however we did not respect the hierarchical relations specified by the DDC, as in the case of the ARCHITECTURE domain shown in Table 2. ARCHITECTURE has been mapped onto 720:*Architecture* and TOWN PLANNING onto 710:*Civic & landscape art*.

WDH Domains	DDC Codes
Architecture	[645,690,710,720]
Town Planning	710
Buildings	690
Furniture	645

Table 2: A fragment of WDH for ARCHITECTURE

However, whereas the 710 code is sibling of 720 in the DDC, TOWN PLANNING is child of ARCHITECTURE in WDH. Also, ARCHITECTURE and TOWN PLANNING should be under ART according to the DDC, but they have been placed under APPLIED SCIENCE in WDH.

3.2 Disjunction

This property requires that no DDC code is associated to more than one WDH label. In only one case this requirement has not been met. Apparently, the DDC does not distinguish between the disciplines of Sociology and Anthropology, and reserves the codes that go from 301 to 307 to both of them. Although these two disciplines are strictly connected, it seems to us that in the current practice they are considered as distinct. So the WDH contains two distinct domains for SOCIOLOGY and ANTHROPOLOGY, which partially overlap because they both map onto the same DDC codes 301:307.

3.3 Basic Coverage

The term *basic coverage* refers to the ideal requirement that all human knowledge be covered by the totality of the *Basic Domains* (i.e. the domains composing the second level of WDH). Also in this case, we used the DDC as a gold standard to measure the coverage of WDH. Given the fact that the DDC has been used for more than a century to classify books and written documents all over the world, we can assume that the DDC guarantees a complete representation of all branches of knowledge. So the *basic coverage* has been manually checked by verifying that all (or almost all) the DDC categories can be assigned to at least one *Basic Domain*.

From a practical point of view, it would be very complicated to check all the thousands of codes contained in the DDC. Thus, our check relied on two assumptions. First, when the *Basic Domains* are taken as a stand alone set, the semantics of a *Basic Domain* is given by its specific code together with the codes of its subdomains. Second, once a DDC code is covered by a *Basic Domain*, inductively, all the more specific categories are covered as well. These assumptions allowed us to actually check only the topmost DDC codes. For example, let's take the **3**00 main class of the DDC. Table 3 below shows that all the sub-codes of the **3**00 class are covered by one or more domains.

In order to improve the overall WDH coverage, 5 completely new domains have been introduced (the first three are *Basic*): PARANORMAL, HOME, HEALTH, FINANCE and GRAPHIC ARTS.

Codes	DDC Categories	WDH Domains
300	• *Social sciences*	• SOCIAL SCIENCE • SOCIOLOGY • ANTHROPOLOGY
310	• *General statistics*	• SOCIOLOGY
320	• *Political science*	• POLITICS
330	• *Economics*	• ECONOMY
340	• *Law*	• LAW
350	• *Public administration & military service*	• ADMINISTRATION • MILITARY
360	• *Social problems & services*	• SOCIOLOGY • ECONOMY • SEXUALITY
370	• *Education*	• PEDAGOGY
380	• *Commerce, communication, transport*	• COMMERCE • TELECOMMUNICATION • TRANSPORT
390	• *Customs, etiquette, folklore*	• FASHION • ANTHROPOLOGY • SEXUALITY

Table 3: Coverage of the **3**00 DDC class

We can now assume that the domain-coverage of the new version of WDH is almost equivalent to that of the DDC, thus ensuring the complete representation of all branches of knowledge.

The new WDH allowed us to fix a number of synset classifications that were unsatisfactory in the previous version of WORDNET DOMAINS. For instance, in the first version of WORDNET DOMAINS the English/Italian synset {microwave oven, microwave}/{forno a microonde, microonde} was annotated with the FURNITURE domain, while the synset {detergent}/{detersivo} was annotated with FACTOTUM (i.e. no specific domain) as no better solution was available. The new WDH hierarchy allows for a more appropriate classification of both synsets within the new HOME domain.

A few DDC codes are not covered by the new list of domains either. These are the codes under the 000:*Generalities* class which includes disciplines such as 010:*Bibliography*, 020:*Library & information sciences*, 030:*Encyclopedic works*, 080:*General collections*. This section has been specifically created for cataloguing general and encyclopedic works and collections. So it is a idiosyncratic category which is not based on subject but on the genre of texts.

Another set of codes which remains not covered by WDH are those going from 420 to 490 and from 810 to 890. These DDC codes are devoted to specific languages and literatures of different countries, for example, 430:*Germanic Languages*, 440:*Romance Languages*, 810:*American Literature in English*, etc. These codes are undoubtedly relevant for the classification of books, but are not compatible with the rationale of WDH, which is meant to be a language-independent resource.

3.4 Basic Balancing

The requirement about *basic balancing* is meant to assure that all *Basic Domains* have a comparable degree of granularity.

Defining a granularity metrics for domains is a complex issue, for which only a tentative solution is provided here. At a first glance, three aspects could be taken into consideration: the number of publications about a domain, the number of sub-codes in the DDC, and the relevance of a domain in the social life.

As a first attempt, balancing could be evaluated referring to the number of publications classified under each *Basic Domain*. In fact, data are available about the number of texts classified under each of the DDC codes. Unfortunately, the number of books published under a certain category may not be indicative of its social relevance: very specialized domains may include a high number of publications, which however circulate in a restricted circle, with low social impact. For example, the number of texts classified in the History domain turns out to be more then ten times the number of texts catalogued under the Computer Science domain. However, if one looks at the number of HTML pages available on the Internet, or the number of magazines sold in a newspaper stand, or the number of terms used in everyday life, one cannot maintain that History is ten times more relevant than Computer Science.

Another approach for evaluating the granularity of domains could be to take into account the number of DDC sub-codes corresponding to each *Basic Domain*. Unfortunately, also this approach gives results which are far from being satisfactory. The fact that a discipline has many subdivisions seems not to be clearly correlated with its relevance. For instance in the DDC manual (version 21) 105 pages can be put in correspondence with the ENGINEERING domain, whereas only 26 correspond to SPORT. It should also be said that there is no correlation between the number of publications and the number of sub-categories in the DDC. For instance, ARCHITECTURE has a great number of publications classified under it, but on the contrary, the number of sub-categories in the DDC is very limited.

The third criterion to evaluate the granularity of domains is their social relevance, which seems not to be captured adequately by the previous two criteria. Of course, social relevance is very difficult to evaluate. We tentatively took into consideration the organization of Internet hierarchies such as the Google and Yahoo directories, which seem to be closer than the DDC to represent the current social relevance of certain domains. See for instance the huge number of HTML pages classified in Google under the topic *Television Programs*. Of course Internet is only a partial view of the organization of human knowledge, so we cannot simply rely on the Internet to evaluate the granularity of the domains.

None of the approaches analyzed so far seems to fit our needs. Thus we took into consideration a fourth criterion, which is based on the DDC as well. Instead of counting the number of subdivisions under a certain DDC code, we measured the depth of the code from the top of the hierarchy. For instance we can say that 700:*Art* has depth 1, 780:*Music* has depth 2, 782:*Vocal Music* has depth 3, and so on. We make the assumption that two DDC codes with the same depth have the same granularity. For instance we assume that 782:*Vocal Music* and 382:*Foreign Trade* have the same granularity (both have depth 3).

In order to evaluate the granularity of the *Basic Domains* against the DDC, we can compare WDH labels and DDC codes with the same depth. Given that the *Basic Domains* have depth 2, we should compare them to the so called *Hundred Divisions* (000, 010, 020, 030, …, 100, 110, 120, etc.). Summing up, we will say that the *Basic Domains* are balanced if they can all be mapped onto the *Hundred Divisions*. Also, in the comparison we should take into account that the *Basic Domains* are 45, whereas the *Hundred Divisions* are 100. So, we expect that in the average, one *Basic Domain* maps onto two *Hundred Divisions* with a small degree of variance with respect to the average.

What we have obtained from the analysis of the new WDH is the following: out of 45 *Basic Domains*

- 4 domains map onto a *Main Class* (depth 1)
- 18 domains are mapped at the *Hundred Divisions* level (depth 2)
- 6 domains are mapped at different DDC levels, with the majority of DDC codes at depth 2
- 17 domains map onto subdivisions of depth 3 and 4.

As for the average number of DDC codes covered by each *Basic Domain*, the variance is quite high. Certain *Basic Domains* cover a big number of codes from the *Hundred Divisions*. For instance HISTORY, and ART cover 6 codes each. Instead, in most cases, one *Basic Domain* covers only one DDC code (e.g. LAW and 340:*Law*).

The evaluation of the granularity of the *Basic Domains* according to the proposed criterion can be considered satisfactory even if the results diverge somewhat from what expected in principle.

To explain this partial divergence in the granularity of domains, one should take into

consideration that the DDC has been created relying heavily on the academic organization of knowledge disciplines. On the other side, in the practical WDH reorganization process we tried to balance somehow this discipline-oriented approach, by taking into account also the social relevance of domains. This has been done by relying on the organization of Internet directories and on our personal intuitions.

Such an approach led us to put at the *Basic* level WDH labels corresponding to DDC codes with depth higher than 2 (more specific than the *Hundreds Divisions*). See for instance the positioning of RADIO+TV, FOOD, HEALTH, and ENVIRONMENT at the *Basic* level, even if they correspond to DDC codes of level 3 and 4. Instead, ANIMALS and PLANTS were not *Basic* in the previous version of WDH, but have been promoted to the *Basic* level in accordance with the granularity level they have in the DDC.

Other domain labels have been placed at a lower level then expected with reference to the DDC. For instance PHILOSOPHY, ART, RELIGION, and LITERATURE have been put at the *Basic* Level, even if they correspond to DDC codes belonging to the *Main Classes* (depth 1). On the other side ASTROLOGY, ARCHAEOLOGY, BODY CARE, and VETERINARY which were *Basic* in the previous version of the WDH, have been demoted at a lower level in accordance with the granularity they have in the DDC. Only in one case this process of demotion has led to the elimination of a sub-domain, that is TEXTILE.

4 Conclusions

In this paper we described the revision of the WORDNET DOMAINS Hierarchy (WDH), with the aim of providing it with a clear semantics, and evaluating the coverage and balancing of a subset of the WDH, called *Basic Domains*. This has been done mostly by relying on the information available in the Dewy Decimal Classification (DDC). A semantics has been provided to the WDH labels by defining one or more pointers to DDC codes. The coverage of the *Basic Domains* has been evaluated by checking that each DDC code is covered by at least one *Basic Domain*. Finally, balancing has been evaluated mostly by comparing the granularity of the *Basic Domains* with the granularity of a subset of the DDC called the *Hundred Divisions*. Balancing is the aspect of the *Basic Domains* which diverges more clearly from the DDC. This is explained by the fact that we took in higher consideration the social relevance of domains.

We think that the new version of the WDH is better suited to act as a useful language-independent resource in the fields of computational lexicography, corpus building, and various NLP applications.

5 Acknowledgements

Thanks to Alfio Gliozzo for his useful comments and suggestions about how to improve the WORDNET DOMAINS Hierarchy.

References

BALKANET http://www.ceid.upatras.gr/Balkanet/

L. Bentivogli, C. Girardi and E. Pianta. 2003. The MEANING Italian Corpus. In *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster, United Kingdom.

C. Fellbaum. 1998. *WordNet. An Electronic Lexical Database*. The MIT Press, Boston.

B. Magnini and G. Cavaglià. 2000. Integrating Subject Field Codes into WordNet. In *Proceedings of LREC-2000*. Athens, Greece.

B. Magnini, C. Strapparava, G. Pezzulo and A. Gliozzo. 2002. The Role of Domain Information in Word Sense Disambiguation. *Journal of Natural Language Engineering (Special Issue on evaluating Word Sense Disambiguation Systems)*, 9(1):359:373.

J.S. Mitchell, J. Beall, W.E. Matthews and G.R. New (eds). 1996. *Dewey Decimal Classification* Edition 21 (DDC 21). Forest Press, Albany, New York.

E. Pianta, L. Bentivogli and C. Girardi. 2002. MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First Global WordNet Conference*. Mysore, India.

G. Rigau, B. Magnini, E. Agirre, P. Vossen and J. Carrol. 2002. MEANING: a Roadmap to Knowledge Technologies. In *Proceedings of the COLING-2002 workshop "A Roadmap for Computational Linguistics"*. Taipei, Taiwan.

H. Schutze. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97-123.

SIMPLE. 2000. *Linguistic Specifications*. Deliverable D2.1, March 2000.

P. Vossen (ed). 1998. *Computers and the Humanities (Special Issue on EuroWordNet)*, 32(2-3).

D.E. Walker and R.A. Amsler. 1986. *Analyzing Language in Restricted Domain. Sublanguage description and Processing*. Lawrence Earlbaum, Hillsdale NJ.

Appendix : The first two levels of the WDH new version with the corresponding DDC codes

TOP-LEVEL	BASIC DOMAINS	DDC
Humanities		
	History	[920:990]
	Linguistics	410
	Literature	[800, 400]
	Philosophy	[100-(130, 150, 176)]
	Psychology	150
	Art	[700-(710, 720, 745.5, 790-(791.43, 792, 793.3))]
	Paranormal	130
	Religion	200
Free_Time		[790-(791.43, 792, 793.3)]
	Radio-Tv	[791.44, 791.45]
	Play	[793.4:795-794.6]
	Sport	[794.6, 796:799]
Applied_Science		600
	Agriculture	[338.1, 630]
	Food	[613.2, 613.3, 641, 642]
	Home	[640-(641, 642, 645)]
	Architecture	[645, 690, 710, 720]
	Computer_Science	[004:006]
	Engineering	620
	Telecommunication	[383, 384]
	Medicine	[610-(611, 612, 613)]
Pure_Science		500
	Astronomy	520
	Biology	[570-577, 611, 612-612.6]
	Animals	590
	Plants	580
	Environment	577
	Chemistry	540
	Earth	[550, 560, 910-(910.4, 910.202)]
	Mathematics	510
	Physics	530
Social_Science		[300.1:300.9]
	Anthropology	[301:307, 395, 398]
	Health	[613-(613.2, 613.3, 613.8, 613.9)]
	Military	[355:359]
	Pedagogy	370
	Publishing	070
	Sociology	[301:319-(305.8, 306.7), 360-(363.4, 368)]
	Artisanship	[338.642, 745.5]
	Commerce	[381, 382]
	Industry	[338-(338.1, 338.642), 660, 670, 680]
	Transport	[385:389]
	Economy	[330-(334, 338), 368, 650]
	Administration	[351:354]
	Law	340
	Politics	320
	Tourism	[910.202, 910.4]
	Fashion	[390-(392.6, 395, 398), 687]
	Sexuality	[155.3, 176, 306.7, 363.4, 392.6, 612.6, 613.96]
	Factotum	