Exploring Hindi WordNet as a Lexical Interface and Subject Headings Tool in Library OPAC

B.A.Sharada Ph.D.,
Librarian
Central Institute of Indian Languages
Manasagangotri, Mysore-570 006, INDIA
sharada@ciil.stpmy.soft.net; sharadaba50@gmail.com

Abstract

The most important technical work of the library is cataloging. Previously, for books in Indian languages, the entries were made in Roman transliteration. The software revolution has enabled the preparation of catalog of books in Indian languages in the script of the original language itself. The library management software should be UNICODE compliant for this. Though the cataloging is done in Indian languages, the information retrieval component, 'Subject heading', has to be rendered in English owing to lack of information retrieval tools such as Classification Schedule, Subject Heading (SH) list, etc. in Indian languages. In order to support information retrieval, WordNet was used for English books in the library OPAC as a lexical interface. Similarly, for Indian languages, the Indo WordNet came handy but was available only for Hindi and Marathi. This paper presents a study of how best SH in Indian language can be derived using Hindi WordNet for the books in Hindi, in addition to tools such as subject dictionaries, glossaries, thesaurus etc. if the terms are not included in the WordNet.

1 Introduction

Information processing and retrieval in Indian languages in the digital environment have faced many challenges. The advent of powerful library management software with Unicode support for Indian languages has helped, to a certain extent, in organizing and retrieving the metadata in the original language except Subject Heading (SH). The major drawback is the non-availability of indexing tools in Indian languages for providing Subject Headings. SH is sharp and equal to summarized text. Among the Indian languages, the present study has selected Hindi as the target language. Though the Hindi thesaurus is available, it is in Roget's thesaurus (Kumar & Kumar 2007) pattern that differs from information retrieval thesaurus which can play a complementary role in information retrieval too. Hence, in order to provide the SH and also use Hindi WordNet (HWN) as a lexical interface in the library OPAC, an exploratory study was done in which the Hindi WordNet (HWN) could be used to enhance information search and retrieval in Hindi. This will aid the lay-user to understand better the terminology and the concepts used in the database. Though some studies have been done on the application of wordNet in information retrieval, only few are available in Indo-wordNet applications.

1.1 Controlled Vocabulary

As any natural language, indexing language (IL) too has vocabulary. The IL is an artificial language used in the tools for information retrieval. Since SH list is a part of IL, this also follows 'Controlled Vocabulary'. Owing to the flexibility of natural languages, since many terms representing the same meaning in the IL environment, one standard term or a descriptor to represent a concept will be used. The same pattern is adopted in preparing the thesaurus as well. The main components of the thesaurus Broader Term (BT) and Narrower Term (NT) will be in controlled vocabulary and the Related Term (RT) will be in a normal mode. For example:

Term : Family
BT : Social Institutions
NT : Birth Order, Parents, Children

RT : Domestic Relations; <u>Households</u>; Kinship; Marriage; Matriarchy; Patriarchy; Adult Children; etc.

In the absence of information retrieval tools in Indian languages such as classification schemes, SH list and information retrieval thesaurus, words used in rendering the title play a vital role in providing the SH. This has to be used without violating the IL rules.

2 Sample for the study

As per the government of India O.M.No.11/20015/21/94-OL (K-2) dated 06.01.1995, it is mandatory to spend 50 % of the amount spent on purchase of Hindi books, excluding the expenditure incurred on journals and standard reference books from the total allocated library grant in any library attached to central government institutions in India. Being the representative of all the languages, the Central Institute of Indian Languages (CIIL) library has more than 60% of its collection in Indian languages and 20% Hindi books. For the present study, two hundred titles in Hindi dealing with disciplines such as Sociology, History, Administration, Linguistics, Folklore, etc., were taken for the analysis. While selecting the titles, due care was taken to select titles dealing with subjects other than Literature that come under the literary forms such as Poetry, Drama, Novel, Short Stories etc. since, in these cases, while rendering the SH, meaning of the title is not considered. In rendering the SH for these titles, same descriptors were followed. For example, the unique SHs for all these titles were 'Hindi Literature', 'Poetry', 'Drama', etc. In case of Collection of Essays, the topics may differ since one book consisting of a collection of essays may deal with a whole range of subjects. Though unique classification number H824.08 was given for Essays, in rendering the SH, the contents of the book could have been highlighted. One advantage in the digital library OPAC is that every component of the MARC tag is a search field. OPAC and MARC are explained in the following section.

3 OPAC

OPAC stands for 'Online Public Access Catalog'. It is a computerized online catalogue of the materials held in a library. The library staff and the public can usually access it at several computer terminals within the library, or from outside via the Internet. Since the mid-1980s, it has replaced the card catalogue in most libraries. Since the mid-1990s dedicated terminal-based OPACs have been gradually replaced by web-based OPACs.
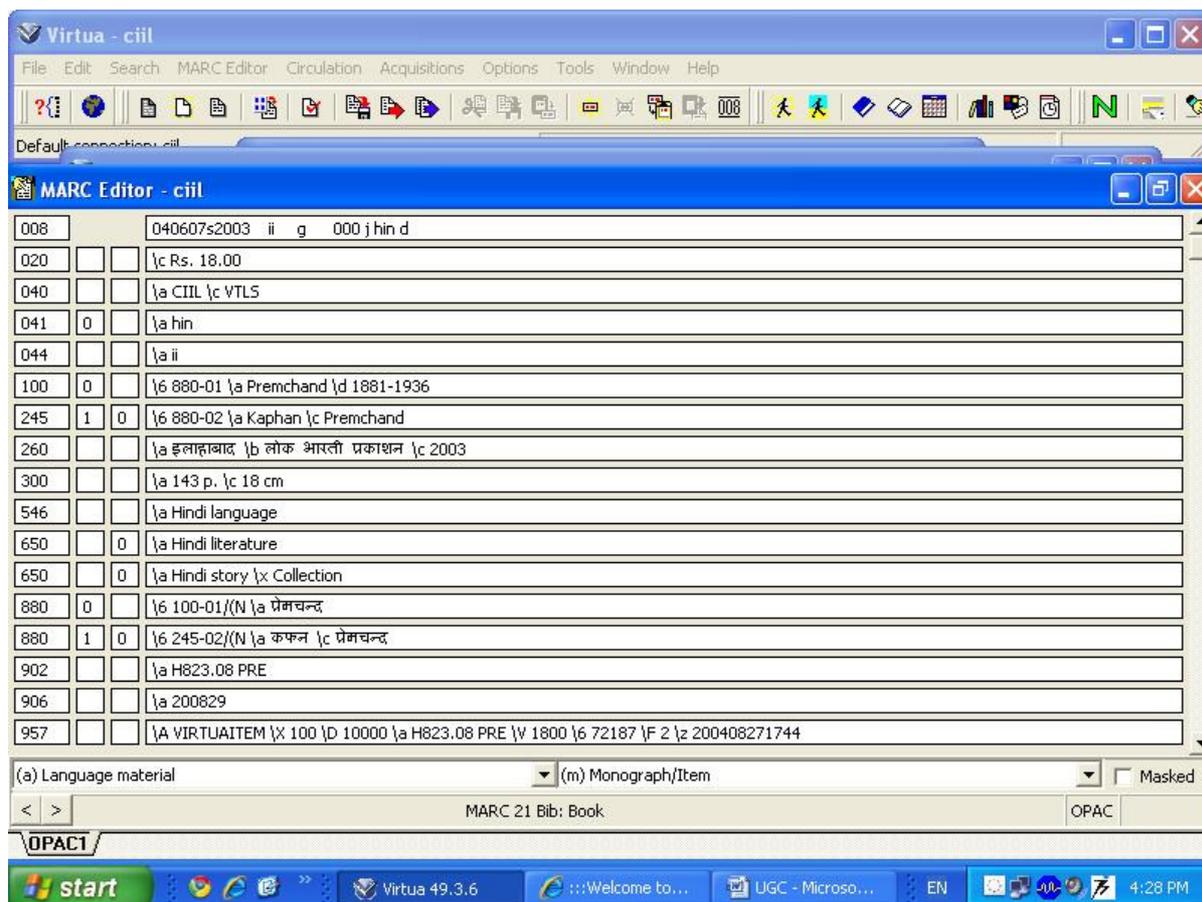
In an OPAC search screen, one can type in the required title, author, subject or keyword in the text box provided and select the icon that fits the search from the dropdown menu. Usually, for research oriented studies, both subject and keyword searches are frequently used. The result is the list of documents available in the library, arranged randomly. If it is a book, a small icon of a book appears. Similarly, different icons appear corresponding to different document forms such as CD's, maps, microfilms, videos, and newspapers. All searches can be stored in a 'cart' in order to help the readers choose multiple results. The cart contains all the bibliographic information including the call number. From the random search list, after a particular result is selected and clicked, three types of catalogues are normally encountered.

- Full
- Items
- Marc

In 'Full' window, the search term is highlighted with full bibliographic information about a document or book. 'Items' window provides the bibliographic information with specific guidelines to enumerative data. In the Marc window, MARC tags related to bibliographic fields along with the data rendered are depicted.

3.1 MARC

The MARC (MAchine Readable Cataloging) format is the international standard for creating computerized bibliographic records. It encodes various descriptive elements of a resource - title, author, physical elements, subjects, etc. into specified fields, each with numerical 3-digit indicators known as tags, that the program recognizes and translates into the data seen onscreen. Most fields are comprised of at least one subfield, which further details the contents of a field (\a); they are represented alphabetically, with each letter having a different meaning with each field. A sample MARC worksheet is given below.

008				040607s2003 ii g 000 j hin d
020				\c Rs. 18.00
040				\a CIIL \c VTLS
041	0			\a hin
044				\a ii
100	0			\6 880-01 \a Premchand \d 1881-1936
245	1	0		\6 880-02 \a Kaphan \c Premchand
260				\a इलाहाबाद \b लोक भारती प्रकाशन \c 2003
300				\a 143 p. \c 18 cm
546				\a Hindi language
650		0		\a Hindi literature
650		0		\a Hindi story \x Collection
880	0			\6 100-01/(N \a प्रेमचन्द
880	1	0		\6 245-02/(N \a कफन \c प्रेमचन्द
902				\a H823.08 PRE
906				\a 200829
957				\A VIRTUAITEM \X 100 \D 10000 \a H823.08 PRE \V 1800 \6 72187 \F 2 \z 200408271744

(a) Language material   (m) Monograph/Item   ☐ Masked

<   >   MARC 21 Bib: Book   OPAC

OPAC1

While cataloging, the language code has to be selected from the 008 tag subfield along with other delimiters for entering the data. For instance, Language code in the above example is 'hin' for Hindi. The 008 tag provides coded information about the record as a whole and a special bibliographic aspect of the item being catalogued. These coded data elements are potentially useful for retrieval and data management systems. In a multilingual context, the bibliographic worksheet, should be designed bilingual - Roman transliteration and original language, and though both the entries have the same data in different language scripts, they are entered in different tags. This helps in data extraction in the required form. Normally, in any of the MARC based library software, English is default and MARC tag is 008, which is a controlled field that enables data retrieval. The centralized database is created in English, which is accessible to the rest of the world. In the above example, in tag 245, the title is rendered in Roman script. The same in Devanagari script is rendered in tag '880'. But the tag '650' meant for rendering subject headings (repeatable field) is in Roman script because of the non-availability of indexing tools in Indian languages.

3.2   Providing SH in Cataloguing

Subject access is crucial to the successful utilization of library catalogs. Subject Headings(SH) are very important in cataloging and very helpful in searches. They are excellent ways to narrow or broaden the search and find all the books on the selected topic. While cataloging books for providing the SH, the first approach corresponds to the SH list such as *Sears List of Subject Headings* or *Library of Congress Subject Headings*(LCSH). The next choice is one of subject dictionaries/glossaries, subject thesaurus etc. WordNet was introduced in 2001; since then it has been used both for rendering SH and as a lexical interface for information retrieval for English books.

With reference to books in non-English language, it is necessary to add headings or subdivisions that are necessary for particular interest to the specific geographic area or the needs of the patrons of a given library. Libraries often have local collections of historical nature, or covering a particular regional interest including the language factor. Materials that fit these collections may need some additional or modified subject headings to meet the needs of

the collection. Here, one feels the need of SH in different languages. Since the approach for books in Indian languages is in the original language and not in English, rendering the SH in English for these books is compromised. Moreover, it is cumbersome both to translate the subject into English, as well as provide the SH depending on the availability of the term in the SH list.

3.3 WordNet and Subject Heading

Providing thoughtful, carefully done subject analysis of materials added to the library collection is an important step in good library service. Different types of SH are discussed in my earlier study (Sharada 2002) including the advantages drawn from wordNet in deriving the SH. Also it was stated that WordNet can be used in conjunction with SH list and domain specific thesauri to improve information retrieval. Because different domains contain similar concepts described with different terminology. Formerly in a card catalogue there existed the limitation of space, where as in the digital environment, one could choose any number of

relevant subject headings in a more user friendly natural language using WordNet in conjunction with subject heading lists and thesauri. Among Indian languages the WordNet is available for Hindi(HWN) and Marathi languages only. In the CIIL Library OPAC these two have been included as a lexical interface for information retrieval. In the present study a trial was made how best HWN could be used to render the SH for books in Hindi, so that SH field also could be in Hindi in the OPAC.

4 Analysis

4.1 Word Frequency Count

Using the 'Corpora Tools' in the 'Indian languages Corpora' prepared by the CIIL (http://www.ciilcorpora.net/frequency.asp) word frequency count was done. Here, the two hundred (200) Hindi titles selected as sample for the study were run in the Word frequency count program. A sorted list of words with their frequency was obtained. The most frequent words are reported in Table 1

Sl.No	Descriptors in Hindi			Frequency	Rank
1	भारत	*bhArata*	India	32	1
2	हिन्दी	*hindI*	Hindi	19	2
3	विज्ञान	*vijnj'Ana*	Science	17	3
4	भाषा	*bhASA*	Language	13	4
5	संस्कृति	*saMskriti*	Culture	11	5
6	इतिहास	*itihAsa*	History	10	6
7	विकास	*vikAsa*	Development	9	7
8	शिक्षा	*shikSA*	Education	9	8
9	लोक	*lOka*	Folk	8	9
10	जीवन	*jIvana*	Life	6	10
11	सिद्धान्त	*siddhAnta*	Theory	6	11
12	साहित्य	*sAhitya*	Literature	5	12
13	कला	*kalA*	Art	5	13
14	समस्या	*samasyA*	Problem	5	14

Table 1- Word Frequency Count

In rendering two hundred titles in Hindi, the authors have used 886 words. In order to check

the most frequently used words, the close category form words such as 'ka', 'ke', 'aur',

'se', etc. have not been taken into consideration. The content words give an interesting result as depicted in Table-1. This table is prepared limiting the word frequency to 5 times each. Other than the words mentioned in the above table, 24 words have been used 4 times, 16 words 3 times, 57 words 2 times and 339 words single time. This helped in noting down the meanings from different sources.

4.2 Expressive Title

The title of the document is expected to reflect the complete content. It may be rendered in a single word or a phrase. While naming a document, the author focuses more on semantics rather than the syntactic factors, since proper concepts have to be selected to represent the complete content of the document irrespective of any language. Brooks, BC (1968) stated that, ranking of terms/group of terms offers an inefficient, precarious basis for indexing and retrieval system. But in Hindi, it is very much direct in concept representation compared to that in English. An important observation while analyzing the Hindi document titles was, they were too expressive. The advantage of expressive title is, it is enough if the content words/concepts or the subject expressive words are indexed. In the Indian language situation, wherein IL tools are not available, if the titles are expressive, these expressions could be used as they are. Since the IL uses controlled vocabulary,

these terms have to be used in such a way that vocabulary of IL is not disturbed.
Example: देवनागरी लिपी तथा हिन्दी वर्तनी का माननीकरण *dEvanAgarI lipI tathA hindI vartanI kA mAnanIkaraNa* - Standerdization of Devanagari script and Hindi spelling
SH would be: bhASA shAstr (linguistics), *dEvanAgarI lipI* (Devanagari Script), *hindI* (Hindi), *vartanI* (Spelling), *mAnanIkaraN* (Standerdization).

Usually the concepts will be a single word representation. But in many cases the semantic representation cannot be understood if the related words are not there, for example, आर्थिक विकास *Arthika vikAsa* 'Economic development'.

5 Hindi WordNet (HWN) as SH tool and Lexical interface

Before any new thing is being adopted to the conventional system, one has to check its appropriateness. So is the case in adopting HWN into the OPAC.

5.1 Discussion

In order to use the HWN as an auxiliary tool to search appropriate SH in Hindi for each title, all the terms present in the title were checked in HWN. For illustration, analysis of five titles is presented in Table - 2.

	a. Title of the Document	b. SH suggested by Hindi expert and Thesaurus	c. SH in the LCSH	d. SH found in HWN
1	आर्थिक विकास का केन्द्रीय सिद्धांत Central theory for economic development	*Arthika vikAsa*, (Economic development) *Arthika siddhAMta* (Economic theory) केन्द्रीय सिद्धांत *kEndrIya siddhAMta* (Central theory)	Economic development Economic Policy Economics Industrialization Central theory - Not available (NA)	आर्थिक(economic) वित्तीय (financial) रुपये-पैसे (Rupees-Paise currency) विकास उन्नति, (development) उत्थान तरक्की प्रगति (development) सिद्धान्त (theory) नियम (Rules)
2	भारत की आदिवासी महिलाएं Tribal women of India	भारत *bhArata* (India), आदिवासी *AdivAsI* (Tribal), आदिवासी-महिलाएं	India Tribes Women Tribal women (NA)	भारत, हिंदुस्तान, इंडिया (India) आदिवासी, मूल-निवासी (Tribal)

		AdivAsI mahilAEM (Tribal women)		महिला, स्त्री, औरत, नारी (women)
3	कार्यालय कार्य बोध (work guide to government offices)	कार्यालय *kAryAlaya* (office), कार्य बोध *kArya bOdha* (teaching of work), कार्यालय कार्य *kAryAlaya kArya* (Office work)	NA	कार्यालय, दफ़्तर (Office) कार्य, काम, कर्म, ड्यूटी (Official duties) बोध, संज्ञान, ज्ञान (Teaching)
4	भारत का राजनीतिक संकट (political crisis of India)	भारत *bhArata* (India), राजनीति *rAjanIti* (Politics), राजनीतिक *rAjanItika* (political) संकट *saMkaTa* (crisis)	India political	भारत, हिंदुस्तान, इंडिया (India) राजनीतिक((political), राजनीति विषयक (Subject of politics) संकट, आफ़त, मुसीबत, विपत्ति, (crisis)
5	नाभिकीय भौतिकी (Nuclear Physics)	नाभिकीय, *nAbhikIya* न्युक्लियर nyukliyar (Nuclear) नाभिकीय भौतिकी *nAbhikIya bhautikI* (Nuclear Physics)	Physics Nuclear Physics	भौतिकी, भौतिकशास्त्र (Physics), पदार्थ विज्ञान (Material Science)

Table - 2   SH in Hindi

The terms that are in bold in column b are taken from the Penguin English-Hindi /Hindi-English Thesaurus (Kumar, Arvind, and Kusum Kumar. 2007). It may be observed in the first title, the concept 'Economic Development'. Column 'b' provides appropriate SH, where as in HWN, the concept has to be split as 'Economic' for which we get three meanings and the word 'Development' gets five meanings. Though we get a clear idea about a particular word in HWN it fails in providing the meaning to the total concept which is a very important factor in providing SH. Also one has to be choosy in selecting the appropriate meaning before selecting the same as SH. For example while checking the term 'Army', four meanings are listed in which दल (*dala*), and फौज (phauja) are correct. But the same यूथ (yUtha) and संरक्षण (saMrakSaNa) are not suitable. This aspect related to compound words in English has been discussed in an earlier study (Sharada 2004). Also for the terms available only in Indian context, for example 'पंचायती राज्य' *paMcAyatI rAjya,* English equivalent is not available in LCSH. Though the term *paMcAyat* is available in HWN, the compound word *paMcAyatI rAjya* is not available.

However out of the two hundred titles selected for the study, approximately 70% terms were available in the HWN as a single term representation. From the above table 2, it is observed that, it is not possible to provide the SH based on either LCSH or HWN. Dependence on experts in Hindi as well as subjects and reference tools in Hindi such as glossary, dictionary and thesaurus are very much essential. For example

term नाभिकीय *nAbhikIya* 'nuclear' is not available in HWN but present in LCSH. The same for Sl.No.3 in the above Table 2, the SH are available in HWN but not in LCSH. The exact translation from Hindi to English is too crucial in order to get the exact SH in LCSH.

Even if cataloging is purchased or copied from other sources, subject headings should already be listed in the cataloging record. Often, these cataloging records are copied from MARC records provided by the Library of Congress. This usually ensures that the subject headings are accurately done and usable as they are. Cataloging staff, however, should not assume that all subject headings can just be copied without checking them. For example *Adhunika Hindi kaviyoṃ ke śabda-prayoga* (The usage of words by modern Hindi poets). The SH for this title which will be found in English only are: Hindi language/usage, Hindi poetry, History & Criticism. Problem with this type of cataloging is, the entry is not found in the original language and the English translation is not there. For example: a. *Hindi bhasha aura Devanagari lipi,* b. *Bhasha vijnana aura Hindi bhasha,* c. *Hindi Bhasha aura Sampreshana Prakriya.* The transliteration have different schemes and in some transliteration, capital letters have different value and will not be uniform. Hence it is not possible to adopt this type of cataloging.

5.2 Advantages of HWN

No need to depend upon Roman script for SH

Each synsets chosen for SH will have the link files.

Link files will be with the HWN

In case of doubts with meaning of the words users will have direct access.

5.3 Disadvantage

In HWN single term representation is present.

In many cases of document titles, one has to take the compound word to consideration.

HWN has still to be comprehensive in its coverage. For example, 'खड़ी बोली विकास के आरम्भिक चरण *khaD'I bOlI vikAsa kE Arambhika caraNa* (The beginning of the development of Khadi-boli dialect). Meaning in thesaurus बोली- उपभाषा(dialect), स्थानीय भाषा(Regional language) and खड़ी बोली- कुरु भाषा, तथा हरियाणा की बोली (Khadi-boli: a dialect in the area of Hariyana). But in the HWN

five meanings for 'development' - विकास (vikaas), उत्थान (utthaan), तरक्की (tarakki), प्रगति (pragathi), अभ्युदय (abhyuday) are available and the important dialect is missing.

6 Conclusion

Based on the results, it is better to use the HWN and provide the SH in Hindi instead of rendering in English. It will be nearer to the user's approach. Preliminary work has to be done at the database creation level. Each word from the title has to be entered in the online HWN. If these words are available in HWN, those terms could be rendered as SH in the database with the link files. In seeking the appropriate meaning such link files for each word will be useful to the library staff as well as the readers. The user can have access to terms occurring in different context and domains with the result they can search across domain boundaries imposed by the classification. However before providing the link, appropriate words have to be selected. In the library OPAC, since it is included as lexical interface, information retrieval will be easy and even the user will have direct access. This will clarify their idea, meanings of the term and address other vocabulary issues. Also this will help to develop a pre-search modal for user assistance.

If the HWN is made more comprehensive, and includes compound words also, in the absence of information tools in Indian languages, one can completely depend upon it both for rendering the SHs and as a lexical data base to information retrieval. Now the dependency for SH are LCSH, thesaurus, many subject dictionaries in Hindi listed in the reference (Items 2,5,6,7).

It will also save the time for the professionals and at the same time help the readers in getting the pinpointed information in the original language of the document.

References

Brooks, B.C.1988 Stability of keywords in text of radiological report.

De Costa, Joseph, and Ramshankar Shukla, comps. 1988. *Authentic English-Hindi Dictionary.* Publications India, New Delhi

Finkelstein, Lev, et al. Placing search context. *ACM Transactions on Information Systems.* Vol. 20(1); 2002.

http://www.cfilt.iitb.ac.in/WordNet/webhwn/wn.php

Kumar, Arvind, and Kusum Kumar. 2007. *The Penguin English-Hindi /Hindi-English Thesaurus and Dictionary*. Penguin Books, New Delhi.

Kumar, Arvind, and Kusum Kumar.1996. *samaantar kosh, anukram khand*. National Book Trust, New Delhi

Library of Congress Subject Headings. 17th Edition. Library of Congress, 1993

Sharada,B.A. 2002. *Subject heading in Cataloguing and WordNet*. In proceedings of the National Seminar on Cataloguing Digital Resources. Paper G 1-14.

Sharada,B.A. and Girish,P.M.2004. *WordNet has no 'Recycle Bin'*. Proceedings of the Second International WordNet Conference, GWC 2004, Brno, Czech Republic, page no.311-319, 2004.

Tiwari, Bholanath. *Bhasha Vijnaan Kosh*. Jnaanmandal LTD, Varanasi.

Voorhees, Ellen M. *Using WordNet for text retrieval*. In "WordNet" Ed by Christiane Fellbaum. London: MIT Press, 1999, p285-303.