Automatic Thesaurus Generation from Raw Text using Knowledge-Poor Techniques

Gregory Grefenstette

Automatic Thesaurus Generation from Raw Text using Knowledge-Poor Techniques

Gregory Grefenstette
Rank Xerox Research Centre
Grenoble Laboratory
6 chemin de maupertuis
38240 Meylan, France

Abstract

In addition to showing how lexical units are related within a field, domain-specific thesauri give an idea of what subjects are important to that field and are thus useful at many points in an information system. The major impediment to creation of thesauri has been the cost of their manual creation. We present here a number of automatic techniques that jointly produce a first draft of a thesaurus from any domain-defining collection of text. The techniques are knowledge-poor in that no domain knowledge is required for their use. We have successfully applied these techniques to over twenty corpora ranging from 1 to 6 megabytes. Results from the thesaurus produced from a collection of medical abstracts will also be presented here.

1 Introduction

A large number of software packages (Milstead, 1990) exist that permit the manual creation of thesauri for use as querying or browsing interfaces to textual information sources. The existence of a domain specific thesaurus serves the double purpose of presenting a hierarchical view of the important concepts in the domain, as well as suggesting alternative words and phrases that may be used to describe the same concept in the domain. Knowing alternative ways of expressing a request is a generic problem in computer-based retrieval (Furnas et al., 1987). A thesaurus is therefore considered a valuable adjunct for exploiting the information in an existing document collection. There are however two major problems with the manual creation of domain-specific thesauri. The first problem is the manual effort of identifying the terms to go into the thesaurus and then organizing them. The second problem is the fit or coverage of the manually-created thesaurus to the document collection. After the initial creation, these same problems reappear whenever the document collection, and subsequently the thesaurus, is updated. In response to these problems, much research has examined the possibilities of accelerating manual thesaurus creation or of automatically creating and updating thesauri.

Both problems of identifying important terms and fitting the thesaurus to the document collection can be addressed by computer-based tools which exploit machine-readable corpora. Simple aids to manual construction such as key-word-in-context concordance producers (Salton, 1989) can provide clues to word use in a corpus. Frequency counts of strings have been used (Frakes and Baeza-Yates, 1992)[pp. 173–218] as a semi-automatic method for organizing the vocabulary present in a corpus into classes of general and specific words. Other automatic methods of exploiting the contexts of words by machine comparison have been proposed. On the individual word level, Phillips (1985) used a textual window of five words before and after each word as that word's context. Collecting all this information throughout a corpus allowed statistical clustering of words into equivalence classes. Church et al. (1991) showed that partial syntactic analysis extracting subject–verb–object tuples can be used as similarity-producing context. Using the statistical measure of mutual information, similar words were extracted from comparing these tuples. Ruge (1991) showed that noun phrases can likewise be used as context to recognize similar words. Words modifying the same head nouns clustered into equivalence classes. Besides reducing human construction costs, these automated methods which recognize potentially similar words, to be entered in

cancer :: [255 contexts, frequency rank: 29] MED *Relat.* lesion, tumor; tissue, disease; carcinoma. *Vbs.* advance, disseminate. *Exp.* cancer patient (cf. survival time, joint deformity), cancer chemotherapy (cf. survival time, intra-arterial infusion), cancer cell (cf. human cell, year period). *Fam.* cancer-specific, cancerous.

Figure 1: Automatically extracted thesaurus entries for "cancer" from MED.

a thesaurus, also provide a good fit to the actual document collection since they exploit the same data contained in this collection.

The thesaurus construction tools cited above deal with individual words. Much of a domain dependent terminology also appears as multi-word terms. Research has also been undertaken to identify important multi-word terminology. Evans (1991) scored noun-phrases that were important in a corpus by combining the frequency of each expression, the rarity of individual words composing them, and the coverage of the expression over more specific expressions. Smadja and McKeown (1990) first collects word pairs with strong mutual information, then extracts the contexts around these word pairs to identify fixed domain terminology. These types of methods attempt to identify the important terminology in a collection-defined domain. As an isolated attempt at organizing multi-word terms into some hierarchical structure, Hearst (1992) used lexical semantic patterns to recognize hypernym relations. These patterns provide reliable relations but they are rare and do not necessarily provide information on the most characteristic domain terms.

The above research provides indications that many steps in the creation of a corpus-derived thesaurus may be automated, or at least accelerated. We have been exploring a combination of these types of techniques on large corpora. In this article we will show how a first-draft thesaurus can be automatically constructed from raw text. We have successfully applied these techniques to over 20 corpora, ranging from 1 to 6 megabytes. Some sample results will be presented.

2 Automatic Thesaurus Creation

Figure 1 presents a thesaurus "entry" automatically extracted from a 1 megabyte raw-text corpus of medical abstracts[1] using knowledge-poor techniques. The techniques are knowledge-poor in that no domain knowledge is required for their use. Below we will explain what techniques were used to produce each part of the entry, but first we describe the structure of the entry. The structure of this entry shows that, in this corpus, the word *cancer* possessed 255 contexts (attributes) which were used to judge its similarity to other words. This word *cancer* had the 29th greatest frequency for all the words compared in the corpus, placing it near the top. The name of the corpus use to generate the entry was MED. The entry is then divided into four parts: related words, *relat.*; commonly associated verbs, *vbs.*; common expressions, *exp.*; and words in the same family, *fam.* The different knowledge-poor techniques which are used to extract the information present in each part will now be presented.

2.1 Related words

The related words are extracted by a package called SEXTANT (Grefenstette, 1994). The raw text of a corpus is first tokenized by a regular grammar. The tokens are morphologically analyzed and searched in a lexicon providing the parts of speech of each token. The tagged text is disambiguated by a stochastic disambiguator (de Marcken, 1990) to provide a single tag for each token[2]. The disambiguated text is parsed and the following dependencies between words are extracted and collected. For nouns, all modifying adjectives, verbs of which they are subjects or objects, and other nouns modifying them as appositives or prepositional clauses, are collected as attributes. Similarity between nouns is calculated using a weighted Jaccard measurement. Details of all these parsing steps and similarity results can be found in the previously mentioned reference.

For each noun in the corpus, the result of these comparisons is a ranked list of nouns used in the most similar way to it. To prune this

[1] This corpus is a standard information retrieval testbed available via ftp in the file /pub/med/smart/med.all.Z at ftp.cs.cornell.edu.

[2] The morphological analyzer, lexicon, and statistical disambiguator are not properly part of the SEXTANT package. They were developed at the Laboratory for Computational Linguistics directed by David A. Evans at Carnegie Mellon University, with whose permission they were used in this research.

list, only those words are retained for the entry which are 'reciprocally near neighbors.' This concept is a slight extension of an idea described in (Hindle, 1990). We define $word_1$ and $word_2$ as reciprocally near neighbors if $word_1$ is discovered as one of the N most similar words to $word_2$ and $word_2$ is one of the N most similar words to $word_1$. We arbitrarily chose $N = 10$. As indicated in Figure 1, for the word *cancer* in this medical corpus, the reciprocally nearest neighbors discovered by SEXTANT were *carcinoma, tumor, lesion, tissue,* and *disease.* This *Relat.* list of related words is further divided into three groups according to the relative frequencies of the words. Semi-colons separate words appearing with about the same frequency (*lesion, tumor*) from words appearing more frequently (*disease, tissue*), and then from words appearing less frequently (*carcinoma*) than *cancer* in this corpus. This comparative frequency has been used as a means of creating hierarchical relations between words in a corpus (Frakes and Baeza-Yates, 1992), though it is an extremely weak statistic and certainly better can be found.

2.2 Commonly Associated Verbs

The next section marked *Vbs.* lists a result that is simply obtained from the parsed text. This section shows those verbs for which the noun often appears as the subject or the object or as the head of an attached prepositional phrase. These verbs are used as part of the attributes by which noun similarity is calculated, but here any verb that plays this role three or more times is listed. The most frequently linked verbs are listed first, and the list is truncated at twelve verbs. For the *cancer* example only *disseminate* and *advance* satisfy these criteria.

2.3 Common Expressions

Common expressions involving the entry word are defined as two-word noun phrases one of whose members is the entry word. Recognition of the internal structure of longer phrases is problematic (Warren, 1978) so we restricted ourselves to noun phrases that unambiguously consisted of two words, both of which were tagged as nouns. The ten most frequent expressions containing the entry word and appearing three or more times unambiguously in the corpus were retained. For the entry *cancer* three expressions appeared in this medical corpus three or more times: *cancer patient*, *cancer chemotherapy*, and *cancer cell.*

Each expression is followed by two other expressions which are similar to it. This similarity is found in a different way than that explained above for nouns. Since the frequencies of such expressions in a corpus are much lower than those of individual words, we extracted a wider context than the syntactic contexts used for single words. This approach is suggested by the results by experiments (Grefenstette, 1993) which showed that wider contexts produced better results for rare words than the finer-grained syntactic contexts. The contexts that we extracted to judge the similarity of two-word expressions used the entire sentence as a window. All of the other verbs, nouns and adjectives appearing within ten words before or ten words after the noun phrase became its attributes.

Using this context, a weighted Jaccard similarity measure was calculated between the two-word phrases appearing 5 or more times throughout the corpus. Two hundred twenty-two such phrases were found in the MED corpus. As seen in the entry under the word *cancer*, the expressions which shared the most similar contexts to the expression *cancer patient* were the expressions *survival time* and *joint deformity.* The closest expressions to *cancer chemotherapy* were *survival time* and *intra-arterial infusion*, and the closest expressions to *cancer cell* were *human cell* and *year period.* This last expression, *year period* shows one of the weaknesses of our implementation of knowledge-poor approach: we do not recognize certain recognizable expressions such as dates and measures so that expressions such as *3 year period* are treated the same as *3 human cells.* This is one of many improvements that could be added to the system.

2.4 Family of Words

The fourth section of the entry, labeled *Fam.*, shows the results of a simple process that finds families of words using corpus context and a string matching procedure. During our experimentation (Grefenstette, 1993), we found that using as the context of each word the document numbers that it appears in often groups morphological variants of words together as being close. The same phenomenon appears when sentence numbers are used as context, but to a lesser degree. That variant forms of the same word should appear in the same document is normal, since the same concept may be expressed as a noun, a modifier, or a verb. Such uses imply morphological transformations in English and most natural languages. These morphological variations are numerous and often domain dependent; e.g. medical terminology follows different formation rules than finance. If the morphological rules of the domain have not been sufficiently analyzed and codified, it would be

cancer :: [905 contexts, frequency rank: 16] AIDS *Relat.* disease; failure, ascites, lesion, tumor, carcinoma. *Vbs.* advance, develop, treat, smoke, detect, use, review, randomize, induce, increase. *Exp.* cancer treatment (cf. median follow-up, hospital stay).

Figure 2: Automatically extracted thesaurus entries for "cancer" from AIDS.

interesting to have a procedure for discovering them automatically. We have developed a family discovery program using the document number that word appears in as its context and a heuristic matching algorithm to extract family variants. The heuristic matching compares the first three, four or five letters, depending on word length, of words found to be close using document co-occurrence as a criterion. This algorithm favors non-initial morphological variation [3] although matchings could have been based on more complicated pattern matching such as letter bigram matching (Adamson and Boreham, 1974). Using this scheme over the MED corpus gives word variants such as

abnormality abnormal
acetoacetate acetate
acromegaly acromegalic
adeaminase a-deaminase
adhesiveness adhesion
adrenergic adrenalectomized
alaly alalies
amyloidosis amyloid
bacterium bacterial
bacterium bacteriophage
biliary bile
breakage break
british britain

In the automatically created thesaurus entry for *cancer*, we find the two family words *cancer-specific* and *cancerous*. Again, let us remark that these family words, as any part of this entry, could have been found by hand, but we are displaying what our system produces completely automatically.

3 Corpus Based Dependencies

The word *cancer* appears frequently in one other of our corpora, in AIDS, a collection of recent abstracts on this illness. Producing the entry for *cancer* from this source gives the slightly different result in Figure 2. Here we find the words *lesion, tumor, disease* and *carcinoma* again associated with the word *cancer*. But the word *tissue* becomes less similar possibly denoting a different approach to the disease in this corpus. This different approach becomes clearer in the expressions associated with *cancer*. In this AIDS corpus there seems to be a patient-oriented view with *cancer treatment* being found as related to *median follow-up* and *hospital stay*, whereas in the MED corpus we find a disease-oriented view with *cancer cells*, *joint deformity* and *cancer chemotherapy* being prominent terms. These observations are, of course, only subjective conjectures that the human reader may draw; the SEXTANT system has no knowledge about anything in the world.

It is interesting to compare the automatically extracted entry for the same word from two very different corpora. In Figure 3 we give the entries that our system produces for the word *growth*. The first entry is from MED, the corpus of medical abstracts, and the second is from MERGERS, a corpus of Wall Street Journal articles on mergers; one could guess the sources by the associations made by SEXTANT. The medical corpus generates relations between growths in the physical sense of a growth, while the mergers corpus associates growth with gains, losses and performances. It is also interesting to note that the expression *growth rate* is common in both corpora though associated with very different expressions. In the medical corpus, *growth rate* is associated with *growth retardation*, while in the financial corpus it is associated with *future performance* and *profit margins*.

4 Discussion

We have presented above a few examples of automatically produced thesaurus entries from corpora of raw text. The appendix presents a longer sample from the MED corpus. To produce these entries we have used a wide variety of corpus-derived information. We used fine-grained local lexical-syntactic information to discover word-based similarities; we used a larger sentence-window context for comparing rarer multi-word phrases; and we used presence in an entire document to discover word families. The techniques that we used were primitive from a linguistic stand-point: no semantic markers distinguished nouns, no case frames predicted attachments, no morphological analysis beyond

[3] Corpus dependent errors are generated by not stripping prefixes. For example, in a corpus of AIDS abstracts the two words *transmission* and *transfusion* were found to be in the same family by this heuristic, though they obviously are not.

growth :: [284 contexts, frequency rank: 25] MED *Relat.* tumor; effect, tissue; antigen, protein, development. *Vbs.* retard, stimulate, show, follow, enhance, accelerate. *Exp.* growth hormone (cf. bone marrow, parathyroid hormone), growth rate (cf. growth retardation, folic acid), *Exp.* tumor growth (cf. body growth, tenuazonic acid), growth retardation (cf. dna content, body weight), body growth (cf. tumor growth, body weight). *Fam.*

growth :: [320 contexts, frequency rank: 139] MERGERS *Relat.* level, increase, gain; loss; performance, return, rise, decline, flow, expansion. *Vbs.* say, expect, slow, accelerate, maintain, sustain, forecast, continue. *Exp.* rapid growth (cf. buy-out bid, raise capital), profit growth (cf. electronics group, total revenue), growth rate (cf. profit margin, future performance), growth potential (cf. company spokeswoman, board seat), future growth (cf. specialty chain, bottom line). *Fam.*

Figure 3: Automatically extracted thesaurus entries from two different corpora for the word "growth".

plural forms and conjugated verbs, etc., etc. But the results provide a plausible first-draft of a thesaurus, and give a pretty clear picture of what types of things are discussed in the corpus from which they are extracted. It still remains to be seen what a real thesaurus builder would say, what additional pieces of information that he or she would like, and whether such information might be automatically extracted from a corpus. We suspect that such considerations would depend on the end-use of the thesaurus to be built.

What is presented here is not to be considered as a finished product. Many pieces of information that might be considered important are missing. The techniques used all contain thresholds such that rare events are not recognized. Domain terms which would be considered important by a human reader but which only appear once or twice in the corpus are missed. And, of course, concepts which never textually appear in the corpus can not be discovered since no outside information enters into the system. Longer phrases (more than two words) are also not considered although their shorter composing parts are[4]. Only noun phrases are considered; domain-important phrases which are expressed using verbs or other parts of speech are missed. The accuracy of the information degrades as less information is available about the word[5].

We present these results here rather as a demonstration that something which resembles a hand-built thesaurus may be automatically extracted from raw text using available techniques. As parsers improve, as textual pattern recognizers improve, as discourse analysis makes progress, as semantic dictionaries become available, the contexts that may be used to judge word and term similarity will become richer and cleaner and more complete automatic thesauri may be generated.

5 Sample entries from MED

ability :: [76 contexts, frequency rank 118] MED *Relat.* production, capacity; action, fraction, function; present, inability. *Vbs.* base.

abnormality :: [73 contexts, frequency rank 121] MED *Relat.* disturbance, atresia, manifestation; feature, course, disorder; tendency, impairment, anomaly, nature. *Fam.* abnormal.

absence :: [68 contexts, frequency rank 125] MED *Relat.* presence; addition, sibling, ligation.

acid :: [486 contexts, frequency rank 8] MED *Relat.* dna, fraction, hormone, activity, protein. *Vbs.* saturate, transform, mobilize, increase, extract, esterify. *Exp.* amino acid (cf. testosterone propionate, factor viii), tenuazonic acid (cf. tumor growth, vit d), acid phosphatase (cf. enzyme activity, electron microscopy), acid metabolism (cf. mean concentration, folic acid), folic acid (cf. rat kidney, dna content), acid composition (cf. total lipid, blood glucose).

action :: [166 contexts, frequency rank 57]

[4]For example, in this MED corpus there were 547 two-word noun-phrases that appeared 3 or more times, but only 40 of length three, and one of length four.

[5]See in the appendix how the word *property* which is the 121st most frequent word has less satisfying results that the word *rate* which is the 15th most frequent and which possesses more information

MED *Relat.* effect; influence, ability. *Exp.* action potential (cf. time constant, coronary flow)

activity :: [410 contexts, frequency rank 11] MED *Relat.* level, effect; protein, concentration, amount, number. *Vbs.* increase, show, determine, decrease, reduce, inhibit, enhance, contain, alter. *Exp.* enzyme activity (cf. acid phosphatase, testosterone propionate), surface activity (cf. surface tension, inclusion body).

administration :: [156 contexts, frequency rank 62] MED *Relat.* dose; injection, response, treatment, therapy; deficiency, secretion, infusion. *Vbs.* follow, associate.

blood :: [258 contexts, frequency rank 27] MED *Relat.* level; liver, plasma, marrow, value, serum, oxygen, tension. *Vbs.* increase, study, make, find, estimate. *Exp.* blood pressure (cf. oxygen tension, carbon dioxide), blood flow (cf. carbon dioxide, fluid po2), blood volume (cf. stroke volume, flow rate), blood glucose (cf. newborn lamb, ffa level), peripheral blood (cf. thymus cell, bone marrow), cord blood (cf. ffa level, newborn infant), blood pool (cf. age group, blood volume), blood viscosity (cf. blood cell, stress reaction), blood stream (cf. lymphoid cell, electron microscope), blood disease (cf. adult patient, dna molecule).

cancer :: [255 contexts, frequency rank 29] MED *Relat.* lesion, tumor; tissue, disease; carcinoma. *Vbs.* advance, disseminate. *Exp.* breast cancer (cf. stage iv, cancer patient), lung cancer (cf. cell carcinoma, cell line), cancer patient (cf. breast cancer, total estrogen), cancer chemotherapy (cf. survival time, intra arterial infusion), cancer cell (cf. cell carcinoma, human cell).

case :: [572 contexts, frequency rank 5] MED *Relat.* change, study; patient; result, treatment, child, defect, type, disease, lesion. *Vbs.* present, report, occur, find, describe, study, discuss, use, observe, classify, diagnose, analyze. *Exp.* case report (cf. lupus erythematosus, intra arterial infusion), case history (cf. inclusion disease, hypophysectomized rat), index case (cf. cleft palate, nervous system).

cell :: [1156 contexts, frequency rank 1] MED *Relat.* tissue. *Vbs.* label, find, infect, contain, appear, show, nucleate, culture, transfuse, transform, observe, make. *Exp.* lymphoid cell (cf. bone marrow, thymus cell), tumor cell (cf. tissue culture, hela cell), liver cell (cf. bile duct, serum protein), cell line (cf. lung tissue, tissue culture), hela cell (cf. human cell, human lung), cell culture (cf. pleuropneumonia like organism, mycoplasma strain), cell division (cf. dna synthesis, zona glomerulosa), spleen cell (cf. lymph node, tumor cell), cell type (cf. chief cell, parathyroid gland), mast cell (cf. plasma cell, surface tension).

change :: [549 contexts, frequency rank 6] MED *Relat.* study, effect; alteration, disease, pattern, rise, decrease, difference, response, increase. *Vbs.* occur, observe, show, produce, find, result, mark, induce, associate, reveal, relate, note.

characteristic :: [109 contexts, frequency rank 87] MED *Relat.* decrease; infection, type, pattern, difference; course, similarity, adult, feature. *Vbs.* induce.

chemotherapy :: [75 contexts, frequency rank 119] MED *Relat.* drug; therapy, hypothermia; adjunct, exercise, route, chemotherapeutic, hypertension, palliation, radiotherapy. *Vbs.* receive. *Exp.* cancer chemotherapy (cf. survival time, intra-arterial infusion) *Fam.* chemotherapeutic.

child :: [412 contexts, frequency rank 10] MED *Relat.* result, group; case, patient; reaction, year, woman, form, subject, infant. *Vbs.* disturb, show, study, observe, give, bear, report, present, match, find, diagnose, develop. *Fam.* childhood.

clearance :: [66 contexts, frequency rank 127] MED *Relat.* content, ratio, concentration, excretion; urine, reabsorption, permeability, intake.

component :: [107 contexts, frequency rank 89] MED *Relat.* content, synthesis, fraction, antigen; cause, constituent, source, cholesterol, property. *Vbs.* contain, consist. *Exp.* protein component (cf. wuchereria bancrofti, skin reaction)

composition :: [65 contexts, frequency rank 128] MED *Relat.* distribution; metabolism, content; position, fat, phosphatase, se, purity, homogenate. *Exp.* acid composition (cf. total lipid, blood glucose)

. . .

culture :: [208 contexts, frequency rank 37] MED *Relat.* marrow; animal, specimen, antigen, lung, extract, suspension. *Vbs.* infect, isolate. *Exp.* tissue culture (cf. human lung, electron microscopy), cell culture (cf. mycoplasma strain, actinomycin d).

curve :: [104 contexts, frequency rank 92] MED *Relat.* artery; nomogram, ventricle, gradient, pulse, record. *Vbs.* obtain. *Exp.* pressure curve (cf. right ventricle, left ventricle), dilution curve (cf. right ventricle, left ventricle).

damage :: [101 contexts, frequency rank 95] MED *Relat.* hypertrophy; uvr, infiltration, fibrosis, uptake, necrosis. *Vbs.* induce, result. *Exp.* brain damage (cf. childhood schizophrenia, heart rate)

data :: [155 contexts, frequency rank 63] MED *Relat.* evidence, observation, finding; technique, study, result; problem, report, analysis, experience. *Vbs.* obtain, suggest, indicate, present.

day :: [203 contexts, frequency rank 41] MED *Relat.* time; rat, group, patient; yr, year, week, month, hour, hr. *Vbs.* return, follow, occur, reach, maintain, find, carry.

decrease :: [121 contexts, frequency rank 80] MED *Relat.* characteristic, rise; amount, value, concentration, difference, change, increase; fall, reduction. *Vbs.* show, accompany.

. . .

measurement :: [93 contexts, frequency rank 103] MED *Relat.* deficiency, property, analysis, reduction; rise; evaluation, estimation, determination, record, detection.

mechanism :: [136 contexts, frequency rank 72] MED *Relat.* role, process; factor; investigation. *Vbs.* discuss, investigate, explain.

membrane :: [88 contexts, frequency rank 107] MED *Relat.* dirofilaria, endothelium, lamella, granule, infiltration. *Exp.* cell membrane (cf. basement membrane, flow rate), basement membrane (cf. connective tissue, type ii).

metabolism :: [105 contexts, frequency rank 91] MED *Relat.* synthesis, concentration, content; size, depletion, phospholipid, utilization, composition, glucose, mobilization. *Exp.* protein metabolism (cf. zona glomerulosa, folic acid), acid metabolism (cf. mean concentration, folic acid), carbohydrate metabolism (cf. protein metabolism, blood glucose). *Fam.* metabolic.

method :: [298 contexts, frequency rank 23] MED *Relat.* test; mean, procedure, technique. *Vbs.* use, describe, make, modify, improve, show, present, outline, consider, apply.

. . .

procedure :: [154 contexts, frequency rank 64] MED *Relat.* therapy, treatment, technique, method; examination, criterion, operation, surgery. *Vbs.* describe, carry.

process :: [136 contexts, frequency rank 72] MED *Relat.* mechanism; condition, phenomenon, structure. *Vbs.* involve.

production :: [73 contexts, frequency rank 121] MED *Relat.* capacity, output; development, excretion, synthesis, incidence; respiration. *Vbs.* increase.

property :: [85 contexts, frequency rank 109] MED *Relat.* measurement, extract; fraction, component; nature, determinant, enzyme, composition, distribution, constituent.

protein :: [212 contexts, frequency rank 36] MED *Relat.* dna; activity, acid, growth, hormone; molecule, analysis, antigen. *Vbs.* contain. *Exp.* protein metabolism (cf. zona glomerulosa, folic acid), lens protein (cf. acid metabolism, lens epithelium), serum protein (cf. inclusion body, dilution curve), protein fraction (cf. insoluble protein, ionic strength), protein component (cf. wuchereria bancrofti, skin reaction), protein excretion (cf. adult patient, filtration rate), insoluble protein (cf. protein fraction, m urea).

rat :: [331 contexts, frequency rank 22] MED *Relat.* group; rabbit, day, kidney, infant, mice, dog, mouse, animal. *Vbs.* treat, give, expose, determine, study, receive, produce, feed, maintain, fast. *Exp.* rat kidney (cf. folic acid, testosterone propionate)

rate :: [387 contexts, frequency rank 15] MED *Relat.* result, response, increase; effect, change, level; pressure, time, value, concentration. *Vbs.* increase, decrease, find, reduce, induce, follow, determine. *Exp.* growth rate (cf. growth retardation,

folic acid), flow rate (cf. coronary flow, body temperature), survival rate (cf. radiation therapy, survival time), heart rate (cf. fluid pressure, pressure curve), filtration rate (cf. vitamin d, urine volume).

ratio :: [115 contexts, frequency rank 84] MED
Relat. rise; weight, excretion, content, level, concentration; capacity, clearance, glucose. *Vbs.* increase, make, find, decrease.

References

G. Adamson and J. Boreham. 1974. The use of an association measure based on character structure to identify semantically related pairs of words. *Information Storage and Retrieval*, 10:253–260.

K. Church, W. Gale, P. Hanks, and D. Hindle. 1991. Using statistics in lexical analysis. In U. Zernik, editor, *Lexical Acquisition: exploiting on-line resources to build a lexicon*, pages 115–164. Lawrence Erlbaum, Hillsdale, NJ.

Carl G. de Marcken. 1990. Parsing the LOB corpus. In *28th Annual Meeting of the Association for Computational Linguistics*, pages 243–251, Pittsburgh, PA, June 6–9. ACL.

David A. Evans, K. Ginther-Webster, Mary Hart, R. G. Lefferts, and Ira A. Monarch. 1991. Automatic indexing using selective NLP and first-order thesauri. In *RIAO'91*, pages 624–643, Barcelona, April 2–5. CID, Paris.

William B. Frakes and Ricardo Baeza-Yates, editors. 1992. *Information Retrieval: Data Structures and Algorithms*. Prentice Hall, New Jersey.

George W. Furnas, Tomas K. Landauer, L.M. Gomez, and Susan T. Dumais. 1987. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, November.

Gregory Grefenstette. 1993. *Automatic Thesaurus Discovery Via Selective Natural Language Processing: A Corpus Based Approach*. University of Pittsburgh. PhD Thesis.

G. Grefenstette. 1994. SEXTANT: extracting semantics from raw text, implementation details. *Integrated Computer-Aided Engineering*, 6(1). Special Issue on Knowledge Extraction from Text.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*. COLING'92, Nantes, France, July.

D. Hindle. 1990. Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275, Pittsburgh. ACL.

Jessica Milstead. 1990. Thesaurus software packages. In *Proceedings of the 53rd Annual Meeting of the American Society for Information Science*, Toronto, Canada. ASIS.

Martin Phillips. 1985. *Aspects of Text Structure: An investigation of the lexical organization of text*. Elsevier, Amsterdam.

Gerda Ruge. 1991. Experiments on linguistically based term associations. In *RIAO'91*, pages 528–545, Barcelona, April 2–5. CID, Paris.

G. Salton. 1989. *Automatic text processing*. Addison–Wesley, Reading, Mass.

Frank A. Smadja and Kathleen R. McKeown. 1990. Automatically extracting and representing collocations for language generation. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 252–259.

Beatrice Warren, editor. 1978. *Semantic Patterns of Noun-Noun Compounds*. Acta Universitatis Gothoburgensis, Goteborg, Sweden. Gothenburg Studies in English, 41.