Original scientific paper

Use of Corpus Analysis Tools in Medical Corpus Processing

Evelina Miscin
College of Business and Management 'B.A. Krcelic'
V. Novaka 23, Zaprešić, Croatia
evelinamiscin@yahoo.co.uk

Summary

The aim of this paper is to show the use of three corpus analysis tools – TermeX, Simple Concordance and Collocation Extract - in processing of medical corpus to obtain collocations. Each tool had its function and will be dealt with separately. TermeX tool was used for obtaining a list of most frequent nouns and the analysis of their frequency. Simple Concordance tool was used to get concordances and for manual extraction of collocations. Collocation Extract tool was used for extracting remaining collocations with a greater distance between a collocate and a node and for determining Log Likelihood and Mutual Information of these collocations. All the data were helpful in determining the most frequent collocations necessary to improve collocational competence of users of medical English.

Key words: corpus analysis tools, collocations, corpus linguistics, medical corpus

Introduction

This paper deals with the use of corpus analysis tools in medical corpus processing in order to obtain collocations. Three tools were used: TermeX, Simple Concordance and Collocation Extract. Each tool will be described separately together with its role in extraction and processing of collocations.

Firth's explanation of relationship among words is considered the initial theoretical framework. According to Firth (1957), "a collocation is a mode of meaning and the lexical meaning of a word is realised through multiple meanings on various levels" (Firth 1957:192). Other theoreticians define a collocation as an occurrence in which lexical units co-occur with one or more words (Halliday et al. 1964.:33; Ridout & Waldo-Clarke 1970.; Backlund 1973. 1976.; Seaton 1982.; Crystal 1985.:55; Cruse 1986.:40; Zhang 1993.:1). Collocations are very important in communication since their misuse can lead to misunderstanding. There is a great need for the analysis of collocations since they represent a connection between a word on the one hand and a text, on the other hand. However, there are not many dictionaries of collocations, especially those dealing with English for Specific Purposes (ESP).

That is why this research wants to investigate collocations in medical English trying to find out ways of simplifying their extraction from the corpus. So far, there have not been any studies in the field of medical English and most linguists were focused on the research of English for social sciences (e.g. Howarth 1998). Some researchers have dealt with collocations in natural sciences (Thomas 1993, Baker, Francis and Tognini-Bonelli 1993, Pearson 1998). The analysis of collocations requires many authentic texts and that is why a corpus is important. The first corpus used in the statistical analysis of a text was made in 1961 and had only 135,000 words. Two most famous corpora are – Brown's corpus for American English and LOB (Lancaster – Oslo – Bergen) for British English. A corpus is particularly important in the collocation analysis, since it helps in obtaining data that can be statistically processed and is much more reliable way of checking collocations than the one that relies on native speakers (Sinclair 1991, Partington 1998, Hunston 2002, Krishnamurthy 2000). Thanks to computer technologies it started to be easier to make a corpus. All these are general English corpora without specialised vocabulary. A few lexicologists made a research on specialised corpora like Myers (1989), Kretzenbacher (1990), Banks (1994), Salager-Meyer (1992), Williams (1996), Dubois (1997) and Biber, Conrad and Reppen (1998). It is very important to have a corpus to study collocations and determine their frequency. Sinclair (1991), a pioneer of a corpus research thinks that in making a corpus it is very important to establish core vocabulary and wants to show that lexical repetition is very frequent in scientific terminology. Function/structure words are eliminated, thus leaving content/lexical words. The same principle was used in this research. The emphasis was made on upward collocations, i.e. those collocations where *a* is a collocate and *b* is a node (e.g. in a collocation 'donate blood', 'donate' is a collocate and 'blood' is a node). Each of the tools that were used had its function and they were all leading to the same aim – finding the most frequent nouns and their verb collocations that will be used in checking collocational competence of users of medical English and creating glossary of collocations that might prove to be useful to all the users of medical English.

The structure of the paper is as follows: first, the research programme is laid out. Then, methodology is explained which is followed by results and discussion. In the end, the conclusion is given.

Research questions

In order to achieve the aims of the study, the following research questions were formulated:

- What are the most frequent nouns in the corpus extracted by the corpus analysis tools?
- How can corpus analysis tools be used for collocation extraction?
- What is the Log Likelihood/Mutual Information of these collocations?

Methodology

First, the corpus was made from the online version of *Merck's Manual of Medical Information* (http://www.merckmanuals.com). The text of the book was turned into textual files that contain 1,065,181 words. There was a total of twenty five files that correspond to the chapters of the book: 1. Fundamentals; 2. Drugs; 3. Heart and Blood Disorders; 4. Lung and Airway Disorders; 5. Bone, Joint and Muscle Disorders; 6. Brain, Spinal Cord and Nerve Disorders; 7. Mental Health Disorders; 8. Mouth and Dental Disorders; 9. Digestive disorders; 10. Liver and Gallbladder Disorders; 11. Kidney and Urinary Tract Disorders; 12. Disorders of Nutrition and Metabolism; 13. Hormonal Disorders; 14. Blood Disorders; 15. Cancer; 16. Immune Disorders; 17. Infections; 18. Skin disorders; 19. Ear, Nose and Throat Disorders; 20. Eye Disorders; 21. Men's Health Issues; 22. Women's Health Issues; 23. Children's Health Issues; 24. Accidents and Injuries and 25. Special subjects. Next, the files were sent to the Faculty of Electrical Engineering in Zagreb where they were processed by TermeX tool in order to get the most frequent nouns. Simple Concordance tool was downloaded for processing the corpus to obtain concordances and find collocations. The last step was the use of Collocation Extract tool in order to find remaining collocations and determine their Log Likelihood and Mutual Information.

Tools and results

Three analysis tools are explained in more details to clarify their purpose. TermeX tool was used for extraction of the most frequent nouns in the medical corpus. Since it has limitations, i.e. it cannot extract verb collocations, Simple Concordance tool was used for this purpose. This tool is helpful for extracting concordance and finding adjacent collocations (the proximity is usually between one and three items). In order to find collocations in the range of 2 to 5 words, collocations extract was used and kit also established Mutual information and Log Likelihood. It is discussed below in more details.

TermeX

TermeX (http://takelab.fer.hr/TermeX/) (Seljan et al. 2009) is a tool for automatic collocation extraction and terminology lexical construction. The potential collocations are ranked by the strength of lexical associations; fourteen different lexical association measures are provided, based on Pointwise Mutual Information (PMI), Dice and Chi-square. The tool can extract collocations of up to length four. To this end, the standard bigram measures have been extended as proposed by Petrović et al. (2009). Moreover, TermeX tool enables the manual selection of candidate collocations to be included in terminology lexicon, the inspection of concordances of the extracted candidates, and efficient processing of multiple documents. However, its big disadvantage is that it does not extract verb collocations.

For this research, TermeX tool was used to provide most frequent nouns in the corpus and the number of their occurrences which can be seen in the Table 1. The Table 1. gives the list of ten most frequent nouns in the corpus[1]:

Table 1. The list of ten most frequent nouns in the corpus

Noun	Position	Number of occurrences
Blood	17.	6675
Symptoms	25.	4093
Treatment	38.	3059
Drugs	41.	2880
Heart	43.	2764
Disease	48.	2524
Pain	53.	2288
Infection	57.	2212
Skin	59.	2134
Body	61.	2008

The first column shows the noun, the second its position in the corpus and the third how many times it appears in the corpus. From the Table 1 it can be seen that the first most frequent noun, ‘blood’, appears on the 17th place and it appears 6675 times in the corpus. It is followed by ‘symptoms’, which appears on the 25th place and occurs 4093 times in the corpus. This list proved to be important later when the verb collocations of these nouns were determined.

Simple Concordance

Simple Concordance tool (http://www.textworld.com/scp) enables making a list of concordances that make it easier to find collocations. Each file was processed separately for easier text processing. For each word considered, the tool shows a list of concordances from corpus, known as *Key Words in Context (KWIC)*. Figure 1. gives an excerpt of concordances for the noun *blood.*

```
410          is also used to purify blood by removing harmful
410         or excessive numbers of blood cells or platelets in
410  . To be helpful for purifying blood, hemapheresis must
410   the undesirable substance or blood cell faster than the
411        that are used to purify blood are plasmapheresis and
411    , excess numbers of certain blood cells are removed.
411              (an excess of red blood cells), certain types of
411   leukemia (an excess of white blood cells), and
412     large fluid shifts between blood vessels and tissues that
412      and tissues that occur as blood is removed and returned
413  members or friends can donate blood specifically for one
413    the recipient's and donor's blood types and Rh factors are
413       , knowing who donated the blood is comforting, although
413  one from an unrelated person. Blood from a family member is
415   stem cells rather than whole blood. Prior to the donation
415    into the bloodstream. Whole blood is drawn from the donor,
```

[1] The list of other nouns is given in Miščin 2012.

```
415   a machine that separates the blood into its components
415    and returns the rest of the blood to the donor./DONATION
417      process of donating whole blood takes about 1 hour.
417      blood takes about 1 hour. Blood donors must be at least
417  in good health: their pulse, blood pressure, and
417           are measured, and a blood sample is tested to
418        a person from donating blood include hepatitis B or C
418      , poorly controlled high blood pressure, low blood
418      high blood pressure, low blood pressure, anemia, the
418    of hepatitis, and a recent blood transfusion./Generally,
419       are not allowed to give blood more than once every 56
419  practice of paying donors for blood has almost disappeared;
420    that would disqualify them./Blood Typing/Because
421     Typing/Because transfusing blood that does not match the
421      can be dangerous, donated blood is classified by type. A
421           by type. A person's blood type is determined by
421        proteins (Rh factor and blood group antigens A and B)
421   and B) on the surface of red blood cells./The four main
422     blood cells./The four main blood types are A, B, AB, and
422   and O, and for each type the blood is either Rh-positive or
422     , a person with O-negative blood has red blood cells that
422  with O-negative blood has red blood cells that lack both A
422    . A person with AB-positive blood has red blood cells that
422      AB-positive blood has red blood cells that have A and B
422        and the Rh factor. Some blood types are far more
422   than others. The most common blood types in the United
423  anyone can receive type O red blood cells; thus people with
423      ; thus people with type O blood are known as universal
423    donors. People with type AB blood can receive red blood
423  type AB blood can receive red blood cells from any blood
423       red blood cells from any blood type and are thus known
423   recipients. Recipients whose blood is Rh-negative must
423    is Rh-negative must receive blood from Rh-negative donors,
423   donors, but recipients whose blood is Rh-positive may
423     Rh-positive or Rh-negative blood./After a person is
424   is deemed eligible to donate blood, he sits in a reclining
424     the procedure is painless. Blood moves through the needle
424  bag. The actual collection of blood takes only about 10
```

Figure 1. The example of the part of concordances for the noun 'blood'

The collocations of the most frequent nouns were manually extracted. In that case, the most frequent combinations were two or three words away from the node, i.e., the central noun. Such an examination was necessary also to separate one-word nouns (e.g. blood) from the multi-word ones (e.g. blood cell, blood type, blood transfusion, etc.).

Collocation extract

The next tool is Collocation extract (http://collocation-extract.software.informer.com/3.0/), a tool for extracting collocations from a corpus. Collocation extract determines the lexical association word pairs based on the statistical measures of Log Likelihood (LL) and Mutual Information (MI). These measures help in determining collocation fixedness and measure the strength of lexical units and thus they help in distinguishing strong and weak collocations.

MI measures the statistical independence of words x and y by comparing their joint probability against the joint probability under the independence assumption. The use of MI for collocation extraction was first proposed by Church and Hanks (1990). The higher the MI, the higher the probability that words are lexically associated one with another. Log Likelihood is the probability ratio of the occurrence of one collocational component in the presence of another one and the probability that the same collocational component will occur without the other one. Higher LL denotes that the probability that two collocational components occur together is smaller. Collocation extract enables a user to determine the direction for searching collocations, the span, frequency, level of meaning and distance between two words. It was used for determining the connection between LL/MI and collocational competence, since it was expected that better knowledge of collocations was connected with higher LL/MI.

To process the corpus with Collocation extract, we proceed as follows. First, a corpus is chosen that should be in a plain text format and is put in 'File-Save File List'. Then, statistical methods are chosen, in this case, Log-Likelihood and Mutual Information. Next, the span is chosen, which can be from 2 to 5. The number denotes the number of words in which collocations are searched for. For instance, if '2' is chosen, the tool will look for two-word collocations (bigrams) and this was the number chosen here. Then, the direction for searching collocations is chosen. Since the upward collocations were looked for, the left side was chosen. After that, the minimum frequency of collocations was chosen. The lowest frequency of collocations was chosen, i.e. 1. Then, the statistical significance at the level 'p<.005', 'p>.05' or 'all occurrences' is selected. The option 'all occurrences' was selected and the maximum number of collocations that were extracted. The given value of '500' was kept. In searching two-word collocations the distance between two words has to be determined. If '2' is selected, two words are separated by one word. Since the 'Simple Concordance' tool had already extracted collocations with the distance of 3 or even 4, here, the option 1-6 was selected in order to determine other collocations with a greater distance between the members.[2] Table 2. gives the example of a part of the file obtained by such an analysis for the noun 'diagnosis'.

As it can be seen from the Table 2., the collocate, i.e., the word before the researched noun, is in the first column. Articles, adjectives, nouns and verbs occur as collocates, but verbs were the subject of this research. The second column denotes the number of collocate occurrence, while in the third one the researched noun occurs, i.e. 'diagnosis'. The fourth column mentions the number of its occurrence, which is 1712. After this, the frequency of occurrence of this collocation is shown (e.g. 'confirm diagnosis' occurs 133 times) and the last

[2] More detailed instructions on the use of the tool are mentioned in 'Help" of the tool itself.

column shows Log Likelihood. In this way, all the most frequent nouns were analysed and they were given in the table 3.

Table 2. The result of the analysis of 'Collocation Extract' tool for the noun 'diagnosis'

Word1	Freq1	Word2	Freq2	Freq12	ll
symptoms	4052	diagnosis	1712	439	4144.5531
the	63849	diagnosis	1712	597	2751.9236
and	29840	diagnosis	1712	446	2419.0548
confirm	185	diagnosis	1712	133	1850.5791
make	549	diagnosis	1712	66	621.725
confirms	31	diagnosis	1712	28	414.37192
makes	227	diagnosis	1712	39	396.75928
bases	23	diagnosis	1712	9	108.64861
a	24886	diagnosis	1712	58	102.06458
early	517	diagnosis	1712	15	97.100213
definitive	19	diagnosis	1712	7	83.43613
definite	6	diagnosis	1712	5	72.041094
making	242	diagnosis	1712	10	71.815283
establish	38	diagnosis	1712	6	59.819193
confirming	6	diagnosis	1712	4	54.318792
support	170	diagnosis	1712	7	50.208611
suspect	133	diagnosis	1712	6	44.141423
after	2315	diagnosis	1712	12	37.674508
establishing	12	diagnosis	1712	3	32.97642
considers	13	diagnosis	1712	3	32.428022
specific	486	diagnosis	1712	6	28.695834
suggest	98	diagnosis	1712	4	28.612737
preliminary	3	diagnosis	1712	2	27.157059
establishes	4	diagnosis	1712	2	25.431833
precise	47	diagnosis	1712	3	24.189261
prompt	65	diagnosis	1712	3	22.204354
suspects	71	diagnosis	1712	3	21.667737
accurate	71	diagnosis	1712	3	21.667737
suggests	79	diagnosis	1712	3	21.020867
the	63849	diagnosis	1712	8	19.74817
initial	100	diagnosis	1712	3	19.600239
screening	131	diagnosis	1712	3	17.985242

Table 3: The most frequent verb collocations in the corpus (occurring with the most frequent nouns)

COLLOCATION	TIMES IT APPEARS	LL	MI
receive a kidney	3	17.098098	7.5688532
aggravate the injury	1	8.2737904	12.467579
replace the hip	2	12.301151	10.280053
gain weight	18	210.23742	12.827786
establish diagnosis	6	59.819193	10.171127
tolerate pain	2	20.770593	6.3841196
provide relief	48	258.07393	11.071733
pose the risk	21	58.344398	9.9256855
loosen the secretion	5	59.901538	9.9880785
change the bandage	1	21.076902	16.327632
develop a bedsore	2	16.973625	12.581677
relieve pain	108	248.33414	8.4848646
treat the infection	35	44.184953	38.331006
regain consciousness	1	44.825846	12.10363
induce vomiting	5	76.444688	10.512782
produce pain	3	51.985936	5.5569562
suppress inflammation	3	35.032721	9.0880331
undergo dialysis	2	122.05025	12.285744
detect a lump	1	14.268766	9.6039706
impair memory	2	23.424866	9.8576507
abort headaches	3	39.895228	9.8576507
relieve nausea	4	30.481541	6.9119118
tolerate a drug	1	10.62278	4.4610449
catch a cold	1	16.128507	9.1082614
detect a cancer	10	65.871482	5.9893827
cleanse the wound	2	20.29615	12.683775
transmit a disease	17	139.06414	7.4514984
get/develop symptoms	11/47	22.681104/ 201.61944	7.9952898/ 9.9281756
identify antibodies	1	24.756581	9.291458
relieve anxiety	5	9.0133363	7.9225788
precipitate the attack	1	13.334497	10.904165
suppress a cough	4	70.689533	10.66052
cause discomfort	16	55.418659	5.603248
trigger diseases	1	11.883876	9.9288879

Table 3[3]. shows the most frequent upward verb collocations in the corpus which occur with the most frequent nouns, the frequency of their occurrence, their Log Likelihood and Mutual Information. It can be seen that the most frequent collocation is 'relieve pain', which occurs 108 times and also has quite a high Log Likelihood (248.33414), but not so high Mutual Information (8.4848646). The next one is 'provide relief', which occurs 48 times with also high Log Likeli-

[3] This is only a part of the table. The whole table can be seen in Miščin 2012.

hood (258.07393) and low Mutual Information (11.071733). They are followed by 'treat the infection' (35 times) which has a lower LL (44.184953) and a bit higher MI (38.331006).

Conclusion

Processing of any corpus would be a difficult task without computer tools, especially when dealing with collocations. This paper showed the use of three different corpus analysis tools – TermeX, Simple Concordance, and Collocation Extract in processing of the medical corpus. TermeX was useful for making a list of the most frequent nouns and analysing a list of their frequencies. It was shown that the most frequent noun is blood which occurs 6675 times. Simple Concordance proved to be useful in making a list of concordances and finding collocations. Collocation Extract helped in extracting collocations occurring with the most frequent nouns. It was established that the most frequent collocations were 'relieve pain', 'provide relief' and 'treat the infection'. The next step would be to test collocational competence of users of medical English and to establish the connection between the competence and Log Likelihood/Mutual Information. The aim of all these analyses was to analyse the results of automatic extraction results. The future research would aim to consider collocations which would be the most useful for users of medical English and to make a dictionary of medical English collocations.

References

Backlund, U. The collocation of adverbs of degree in English. Doctoral Dissertation, Uppsala University, Uppsala, Sweden, 1973.

Baker, M.; Francis, G. and Tognini-Bonelli, E. (eds.). Text and Technology. Amsterdam: John Benjamins, 1993.

Banks, D. 'Clause Organisation in the Scientific Journal Article'. *Alsed-Lsp Newsletter* (1994). Vol.17/2.4-16

Biber, D.; Conrad, S. and Reppen, R. Corpus Linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press, 1998.

Church, Kenneth Ward; Hanks, Patrick. Word Association Norms, Mutual Information and Lexicography. In *Proceedings of 27th ACL,* 16(1) (1990), 22-29

Cruse, D.A. Lexical Semantics, Cambridge: Cambridge University Press. 1986

Crystal, David. A dictionary of linguistics and phonetics. Oxford: Basil Blackwell Ltd., 1985.

Delač, Davor; Krleža, Zoran; Dalbelo Bašić, Bojana; Šnajder, Jan; Šarić, Frane. TermeX: "A Tool for Collocation Extraction. Lecture Notes" in Computer Science" (Computational Linguistics and Intelligent Text Processing). 5449 (2009); 149-157

Dubois, B.L. The Biomedical Discussion Section in Context. London: Ablex Publishing Corporation, 1997

Dunning, Ted. Accurate Method for the Statistic of Surprise and Coincidence. In *Computational Linguistics,* (1993), 61-74

Firth, J.R. Papers in Linguistics 1934-1951. Oxford: oxford University Press, 1957.

Halliday, M.A.K. et al. The linguistic sciences and language teaching. London: Longman, 1964.

Howarth, P. "The Phraseology of Learners' Academic Writing' in Cowie, A.P. (ed.). Phraseology: Theory, Analysis and Applications. Oxford: Oxford University Press, 1998.

Hunston, Susan. Corpora in Applied Linguistics. Cambridge: Cambridge University Press, 2002.

Kretzenbacher, H.L. Rekapitulation: Textstrategien der Zusammenfassung von Wissenschaftlichen Fachtexten. Tübingen: Gunter Narr Verlag, 1990.
Krishnamurthy, Ramesh. "Collocation from silly ass to lexical sets" in Heffer, C; Sauntson, H. and Fox, G. (eds.) Words in Context: A tribute to John Sinclair on his Retirement. Birmingham: University of Birmingham, 2000.
Miščin, Evelina. Unpublished doctoral thesis 'Glagolske kolokacije u engleskome jeziku'. Osijek, 2012.
Myers, G. 'The Pragmatics of Politeness in Scientific ARticles.' In *Applied Linguistics* (1989). Vol. 10/1:1-35.
Partington, Alan. Patterns and Meanings. Amsterdam: John Benjamins, 1998.
Pearson, J. Terms in Context. Amsterdam: John Benjamins, 1998.
Petrović, Saša, Šnajder, Jan and Dalbelo Bašić, Bojana. "Extending lexical association measures for collocation extraction." *Computer Speech & Language* 24.2 (2010): 383:394.
Ridout, R. & Waldo-Clarke, D. A reference book of English. London: Macmillan, 1970.
Salager-Meyer, F. 'A Text-Typer and Move Analysis Study of Verb Tense and Modality Distribution in Medical English Abstracts.' In *English for Specific Purposes* (1992) Vol.9: 145-159.
Seaton, B. A handbook of English language teaching terms and practice. London: The Macmillan Press Ltd., 1982.
Seljan, Sanja, et al. "Comparative Analysis of Automatic Term and Collocation Extraction." *The Future of Information Sciences: INFuture 2009-Digital Resources and Knowledge Sharing/ Stančić, H. (2009): 219-228.*
Seljan, Sanja; Gašpar, Angelina. First Steps in Term and Collocation Extraction from English-Croatian Corpus // Proceedings of 8th International Conference on Terminology and Artificial Intelligence. Toulouse, France, 2009. http://www.irit.fr/TIA09/thekey/posters/seljan.pdf
Sinclair, John. Corpus, Concordance, Collocation. Oxford: Oxford University Press, 1991.
Thomas, P. 'Choosing Headwords from LSP Collocations for Entry Into A Terminology Data Bank (Term Bank).' In Sonneveld, H.B. and Loening K.L. (eds.) 1993: 46-68.
Williams, I.A. 'A Contextual Study of Lexical Verbs in Two Types of Medical Research Article.' In *English forSpecific Purposes.* (1996). Vol. 15/3:175-198.
Zhang, X. English collocations and their effect on the writing of native and non-native college freshmen. Unpublished Ph.D. thesis, Indiana University of Pennsylvania, 1993.

Web links

http://www.merckmanuals.com
http://takelab.fer.hr/TermeX/
http://www.textworld.com/scp
http://collocation-extract.software.informer.com/3.0/