0008-3194/2011/56–60/$2.00/©JCCA 2011

Quantitative corpus-based analysis of the chiropractic literature – a pilot study

Neil Millar PhD*
Brian S. Budgell DC, PhD†
Alice Kwong, Hons BScKin, CK§

In this pilot study, a collection of peer-reviewed articles from the Journal of the Canadian Chiropractic Association was analyzed by computer to identify the more commonly occurring words and phrases. The results were compared to a reference collection of general English in order to identify the vocabulary which is distinctive of chiropractic. From texts with a combined word count in excess of 280,000, it was possible to identify almost 2,500 words which were over-represented in the chiropractic literature and therefore likely to hold special importance within this domain. Additionally, readability statistics were calculated and suggest that the peer-reviewed chiropractic literature is approximately as challenging to read as that of nursing, public health and midwifery. Certain words widely considered to be of importance to the profession, for example "subluxation and adjustment," were not particularly prevalent in the literature surveyed.
(JCCA 2011; 55(1):56–60)

KEY WORDS: JCCA, chiropractic, corpus, linguistics

Dans le cadre de la présente étude pilote, un ensemble d'articles évalués par les pairs tirés du Journal de l'Association chiropratique canadienne a été analysé par ordinateur afin de déterminer les mots et les phrases les plus communément utilisés. Les résultats ont été comparés à un corpus de référence en anglais général afin d'identifier le vocabulaire spécifique à la chiropratique. À partir de textes ayant un total combiné de mots dépassant les 280 000 mots, il a été possible de déterminer près de 25 000 mots surreprésentés dans la littérature chripratique et qui ont probablement une importance particulière dans ce domaine. Par ailleurs, des statistiques sur la lisibilité ont été calculées et indiquent que la littérature chiropratique évaluée par les pairs est aussi compliquée à lire que celle liée aux soins infirmiers, à la santé publique et à la profession de sage-femme. Certains mots considérés par de nombreuses personnes comme étant importants pour la profession, par exemple « subluxation and adujstment », n'étaient pas particulièrement prévalents dans la littérature en examen.
(JCCA 2011; 55(1):56–60)

MOTS CLÉS : JACC, chiropratique, corpus, linguistique

* Department of Linguistics and English Language, University of Lancaster
† Graduate Education and Research Programmes, Canadian Memorial Chiropractic College
§ Canadian Memorial Chiropractic College
Corresponding Author: Brian Budgell, Canadian Memorial Chiropractic College, 6100 Leslie St., Toronto, Ontario Canada M2H 3J1; tel: (416) 482-2340 ext 151; email: bbudgell@cmcc.ca
Funding: This work was supported by funds from the Canadian Institutes for Health Research and Canadian Memorial Chiropractic College.
Portions of this work were previously presented at "Reconciling Science and Subluxation – a Colloquium" held at CMCC on October 25, 2009 and at ACC-RAC (Association of Chiropractic Colleges – Research Agenda Conference) 2010, March 18–20, 2010.

Introduction

The domain-specific dialects of biomedical and health language include esoteric technical vocabularies as well as conventions of grammar and discourse which distinguish them from root languages such as general English.[1] Characterization of the dialect of a discipline may provide important cultural insights. On a more pragmatic level, identifying a target dialect also permits definition of the language learning burden imposed on students, and therefore, could greatly enhance strategies to impart fluency.[2] Enhanced communicative competence would likely improve education, patient-practitioner cooperation and communications within and between disciplines. Additionally, a quantitative analysis of a dialect shows language as it actually is rather than as we might wish it were. This is particularly valuable to a discipline such as chiropractic where proximity to or distance from other disciplines is an important consideration in the formulation of educational programmes, legislation and health care policy.

Corpus linguistics provides well validated methods to identify distinctive dialects, such as that of chiropractic. The term corpus refers to a large (usually electronic) archive of samples chosen to be representative of a target language.[3] Specialized computer programs permit the analysis of corpora (the plural of corpus) for such quantitative measures as word and phrase frequency, part of speech and even semantic class (meaning of a word or phrase).[4] Previously, the methods of corpus linguistics have been applied to and have discerned distinctive features of the languages of nursing,[5] public health,[6] and midwifery.[7] However, while debate about the meanings of terms such as *subluxation* and *adjustment* is not uncommon in the chiropractic literature (for example, see[8]), no studies have attempted to quantify the lexical and syntactical features of the professional literature, nor to define the extant usages of key words and phrases. Thus the present study was undertaken to determine the lexical and syntactical features of a corpus of modern chiropractic writings.

Methods

A corpus was created by downloading the full texts of 98 articles – editorials, commentaries and research papers – published in the Journal of the Canadian Chiropractic Association from 2005 to 2008. Notices, short announcements and personal profiles were not included in the corpus. Titles, legends, references, acknowledgements and tables were removed from manuscripts, as were figures. Hence, the remaining textual material consisted overwhelmingly of full sentences. The texts were saved as XML files and meta-data markers were inserted by hand to facilitate later analysis on a section by section basis.

The corpus was analyzed using a number of software programmes, including WordSmith Tools V5.0 (Oxford University Press). WordSmith Tools was used to calculate the number of occurrences of each unique word (referred to as a "type" in the jargon of linguists) and compared the relative prevalence of each type to a reference corpus of general English – the New York Times (NYT) corpus. Types (words) which occurred significantly more often in the chiropractic corpus than in the comparison (NYT) corpus (as determined by log-likelihood) were identified as keywords.[9]

Additionally, using the open access tool Vocabprofile,[10,11] each type was classified as belonging to the General Service List (GSL), the 2,000 most common word families in general English,[12] the Academic Word List (AWL), the approximately 570 word families commonly encountered in academic settings,[13] and off-list, that is not belonging to either of the 2 preceding lists. Their absence from the GSL and AWL means that off-list words are more likely to hold special meaning within a target corpus.[14]

The corpus was also analyzed with the readability statistics function of MSWord 2007 to determine average sentence length, prevalence of sentences in the passive voice, Flesch Reading Ease Index and Flesch-Kincaid Grade Level.

Results

The experimental corpus consisted of approximately 280,000 tokens: individual words, letters and numbers, regardless of number of occurrences. The reference corpus of general English (NYT corpus) comprised approximately 3.6 million tokens. Based on a log-likelihood of >15.13[15] and in comparison to the corpus of general English, 2448 types were significantly over-represented ($p < .01$) in the chiropractic corpus. In the language of corpus linguistics, such words are referred to as "keywords."[9] Approximately 74% of the tokens (total collection of words)

were from the General Service List, approximately 11% were from the Academic Word List, and approximately 15% were off-list.

The 10 most prevalent words (tokens) in the chiropractic corpus were *the, of, #, and, to, in, a, is, that* and *for*. In linguistics, such words are known as function words as they aid in sentence construction but do not convey meaning by themselves. The 10 most prevalent content ("meaningful") words (and their % prevalences in the corpus) were *chiropractic (0.71), treatment (0.53), pain (0.52), care (0.35), patient(s) (0.63), health (0.33), evidence (0.31), practice (0.27)*, *study (0.25) and cervical (0.25)*. All of these words were keywords in the sense that their percentage prevalences were significantly higher in the chiropractic literature than in the reference corpus of general English. Other keywords of somewhat lower prevalence included *clinical, manipulation, spine, profession, symptoms, lumbar, research, technique, position* and *injury(ies)*. *Adjust* and words derived from this root had a collective prevalence of 0.05%. *Subluxation* and *subluxations* had a collective prevalence of 0.04%. The types *he*, *him* and *his* occurred approximately 5 times as often as their corresponding female types *she*, *her* and *hers*. The complete list of keywords is posted at http://bmhlinguistics.org/joomla2/chiropractic.

In 12 instances, the word *subluxation*(*s*) collocated with the word *vertebral*. The other common collocation (7 instances) was with the word *chiropractic*. There was only one instance of the phrase *vertebral subluxation complex*. In approximately 40% of instances, the phraseology implied that the meaning of the term *subluxation* was apparent from context or common knowledge. In other instances, there was explicit reference to a specific definition or the need for a definition. Interestingly, in approximately 25% of instances, the reference to subluxation was emotive, politicized and even explicitly disparaging of the term.

Adjust or words derived from it (*adjusting, adjustment* etc.) occurred 113 times in the corpus. There were 117 instances of *mobilize*, or some variation thereof, such as *mobilization*, *mobilizations*, etc. However, there were 405 occurrences of *manipulate* or some variation thereof, and the word *manipulation* was one of the most common keywords in the literature.

The average number of words per sentence was 23.7. The passive voice occurred in 24% of sentences. Overall, the Flesch Reading Ease Index was 29.0 and the Flesch-Kincaid Grade Level was 14.7.

Discussion

The chiropractic corpus created for this study is comparable in size to one previously created for the nursing literature[5] and likely of adequate size to reasonably represent the written language of the modern Canadian chiropractic profession. The written language is, of course, somewhat different from the spoken language used in educational, clinical and professional encounters, and so the results of this study have limited implications. Nonetheless, the outcomes of this exercise are of pragmatic interest to the profession.

Approximately 15% of the words in the chiropractic literature were off-list. That is to say they did not appear in either the General Service List or the Academic Word List. Such words, *subluxation, lumbar* etc., would therefore likely be unfamiliar even to the well-educated reader who did not have specialist knowledge of chiropractic. This is consistent with findings concerning the literature of public health[6] and the literature of midwifery.[7] Furthermore, chiropractic appears to have its own specialized lexicon. Thus, while it shares keywords such as *patients* and *treatment* with other disciplines,[5–7] it also contains its own particular keywords including, of course, *subluxation* and *adjustment*. On the other hand, words which are conceptually important to chiropractic, such as *subluxation* and *adjustment*, are not necessarily highly prevalent in the literature.

As with the languages of nursing[5] and midwifery,[7] in the chiropractic corpus there was a bias in the representation of masculine versus feminine pronouns and possessive adjectives. However, in the instance of chiropractic, the bias is in favour masculine words. Much of the writing in midwifery concerns the experience of the mother, and so it is not surprising that female references abound. In nursing and chiropractic, a proportion of the literature is also introspective, dealing with the respective professions as a whole and with notable individuals within each profession. To the extent that nursing and chiropractic have historically been populated more by women versus men, respectively, any skewing of the balance in masculine and feminine references may be due to the effect of the introspective literature. This hypothesis could be tested by quantifying the contexts

of masculine and feminine words in the respective corpora.

Pertaining to the accessibility of the literature, measures of readability for the chiropractic corpus fell quite close to those of both public health[6] and midwifery.[7] Thus, while the average number of words per sentence was 23.7 for chiropractic, it was 25.8 and 22.4 for public health and midwifery, respectively. The passive voice was used in 24% of sentences in the chiropractic corpus, versus 26% for public health and 29% for midwifery. The passive voice is more prevalent in biomedical literature than in general English and often results in longer and more complex sentence structure.[16] The Flesch Reading Ease index for chiropractic was 29.0 versus 23.2 for public health and 30.7 for midwifery. Flesch Reading Ease is calculated on the basis of word and sentence complexity[17] and is one of the most widely used measures of readability. A higher readability score indicates that text is easier to read and, by implication, easier to understand. The readability indices for this study suggest that the literature of these three disciplines (public health, midwifery and chiropractic) is generally readable to those with an education equivalent to American college graduation.[17] By contrast, the literature of biomedical domains such as clinical microbiology and infectious diseases is much less accessible.[1]

Conclusion

Although concepts such as *subluxation* and *adjustment* may be important within the discipline of chiropractic, the actual terms were not highly prevalent in the literature which we sampled. This may be a particular feature of the Canadian peer-reviewed literature, and so it would be useful to perform a comparison with literature from other sources. Quantitative analysis of the chiropractic corpus also suggests a gender bias in word choice, with over-representation of masculine words. The converse phenomenon, with over-representation of female references has been reported for nursing[5] and for midwifery.[7] In comparison to the literature of other health and biomedical disciplines, that of chiropractic is reasonably accessible.

The findings of these and similar studies could be used in the design of teaching and testing materials, particularly in creating materials which are appropriate to the language of the discipline and the level of education of the target readership. The full data set and search engine on our project web site would also permit authors, reviewers and editors to determine whether a particular turn of phrase is justified by usage.

The degree to which the current results may be extrapolated to other times, settings and professions remains unknown. However, our group is currently applying the same methodology to historical chiropractic literature and to the literature of other groups of manual therapists.

References

1 Budgell B, Miyazaki M, O'Brien M, Perkins R, Tanaka Y. Our Shared Biomedical Language. In: International Medical Education Conference; 2007; Kuala Lumpur: International Medical University; 2007. p. A17.

2 Chung TM, Nation P. Identifying technical vocabulary. System: Intl J Educ Technol Appl Linguist. 2004; 32:251–263.

3 McEnery T, Tono Y, Xiao Z. Corpus-based language studies; an advanced resource book. London, Routledge 2006, p 347.

4 Rayson P. From key words to key semantic domains. International J Corpus Linguistics. 2008; 13:519–549.

5 Budgell B, Miyazaki M, O'Brien M, Perkins R, Tanaka Y. Developing a corpus of the nursing literature: a pilot study. Japan J Nursing Science. 2007; 4:21–25.

6 Millar N, Budgell B. The language of public health – a corpus based analysis. J Public Health. 2008; 16(5):369–374.

7 Chiba Y, Millar N, Budgell B. The language of midwifery and perinatal care. J Jpn Acad Midwif. 2010; 24(1):1–10.

8 Nelson C. The subluxation question. J Chiropractic Humanities. 1997; 7:46–55.

9 Baker P, Hardie A, McEnery T. (2006). A glossary of corpus linguistics, Edinburgh: Edinburgh University Press.

10 Cobb T. Web Vocabprofile [accessed October 2009 from http://www.lextutor.ca/vp/], an adaptation of Heatley & Nation's (1994) Range.

11 Heatley A, Nation P. (1994). Range. Victoria University of Wellington, NZ. [Computer program, available at http://www.vuw.ac.nz/lals/.]

12 West MP. A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology. London, Longmans Green 1953.

13 Coxhead A. A new academic word list TESOL Quarterly. 2000; 34:213–238.

14 Chung TM, Nation P. Technical vocabulary in specialised texts. Read Foreign Lang. 2003; 15:103–116.

15 Rayson P, Berridge D, Francis B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora, Paper presented at the 7th International

Conference on Statistical Analysis of Textual Data, 10–12 March, Louvain-la-Neuve.
16 Salager-Meyer. A text-type and move analysis study of verb tense and modality distribution in medical English abstracts. Engl Specif Purp. 1992; 11:93–113.
17 Flesch R. A new readability yardstick. J Appl Psychol. 1948; 32:221–233.