Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

A study of closure in a nursing textbooks and journals: A corpus based study

Mazura Mastura Muhammad [1,*], Sahandri Gani Hamzah [2], Saifuddin Kumar Bin Abdullah [3], Chan Siang Jack [1]

[1]*Faculty of Languages and Communication, Sultan Idris Education University, Perak, Malaysia*
[2]*Faculty of Human Development and Education, Sultan Idris Education University, Perak, Malaysia*
[3]*Department of Polytechnic, Ministry of Education, Putrajaya, Malaysia*

ARTICLE INFO

Article history:
Received 28 October 2016
Received in revised form
2 September 2016
Accepted 15 September 2016

Keywords:
Sublanguage
Closure
Constrained
Unconstrained
Clinical language

ABSTRACT

The prime aim of the study is to measure the degree of closure of one form of clinical language (specifically nursing textbooks and journals) in order, first, to determine whether these two restricted forms of clinical language can be rightly categorized as a sublanguage; second, to understand better the linguistic features of the language of the nursing domain; and finally, to better understand the nature of sublanguage. In this study, nursing textbook and journal corpora are compared to weather reports and the BNC Sampler. The findings show that none of the linguistic inventories of these corpora approach closure. Investigations conducted on the weather reports show that the corpus approaches closure at many levels. The BNC Sampler, however, behaves exactly as unconstrained language is expected to. The findings show that the nursing textbooks and journals seem to belong in a middle area between highly constrained language and highly unconstrained language. The findings of the study reveal that the idea of a sublanguage is problematic. The original definition of a sublanguage seems to clearly divide sublanguage or constrained language from unconstrained language and placing both as a dichotomy between two discrete categories. However, the findings regarding the nursing textbooks and journals seem to show that there is no explicit or clear-cut boundary that divides constrained language from unconstrained language.

1. Introduction

Various studies have described a highly specialized language as a subset or subsystem of the general language (Harris, 1968), a jargon (Bross, 1972) or a sublanguage (Sager, 1986; DeVille, 2001; McEnery and Wilson, 1996). A sublanguage is defined as the language used by members of a specialized community to meet the specific functional needs of their occupations (Harris, 1968; Bross, 1972; Sager, 1986).

These studies have defined a sublanguage on the basis of features such as specialized subject matter, specific community, the speaker communicative purpose and domain-specific lexis and syntax. McEnery and Wilson (1996), on the other hand, suggested that a sublanguage should be defined based on its degree of closure. McEnery and Wilson (1996) explained that a sublanguage has a high degree of closure at various levels of description. Based on this notion, they develop a method to measure a sublanguage by quantifying various forms of closure. Hence, we can define a sublanguage as any forms of language that have specialized subject matters, utilized by speakers of specific community with similar communicative purposes and possess domain specific lexis and syntax. Closure, on the other hand, is a particular feature of a sublanguage. Since a sublanguage involves forms of language that are specialized, it is closed or finite. Unlike the natural language, which is considered non-finite, a sublanguage is an enumerable set, which can be gathered and counted (McEnery and Wilson, 1996).

One form of language often considered a sublanguage is the clinical language used in healthcare, which is restricted and utilized by specialized groups of people (Spyns, 1996; Friedman et al., 2002; Travers and Haas, 2003). Travers and Haas (2003) explained that clinical language contains specialized terminology and patterns of occurrence and co-occurrence of words in text. This study aims to measure the degree of closure of one

* Corresponding Author.
Email Address: mazura@fbk.upsi.edu.my (M. M. Muhammad)
https://doi.org/10.21833/ijaas.2017.02.017
2313-626X/

form of clinical language (specifically that of the nursing domain), in order, first, to determine whether clinical language can be rightly categorized as a sublanguage; second, to understand better the linguistic features of the language of the nursing domain; and finally, to better understand the nature of sublanguage

2. Problem statement

The original motive for investigating the language of nursing was pedagogic. This study, however, is not in itself a pedagogical study. But the findings of this study may go on to be used in producing a language syllabus that will meet the language needs of nursing students, as learners should be at the heart of any teaching programme, particularly in the field of English for Specific Purpose (ESP).

When one designs a syllabus and materials of an ESP course, three fundamental elements should be taken into consideration – language descriptions, learning theories and needs analysis. Language descriptions are the manner in which the language system is broken down and described for the purpose of enhancing learning. This will then determine what should be included in the language syllabus (Hutchinson and Waters, 1987). The theoretical basis of a teaching methodology is determined by learning theories which help us to understand how the students learn language as well other domains of knowledge (Hutchinson and Waters, 1987). Finally, needs analysis provides answers to the who, why, where and when questions: who will be involved in the process of designing the language course; why the learners need to learn the language; and when and where the learning will occur (Hutchinson and Waters, 1987; McDonough, 1984). Thus before designing a course, understanding these fundamental issues is significant, as the choice and actions of the learners in learning are determined by what is valuable and meaningful to them (Erickson, 1986).

This study sets out to explore the first element – i.e. language description – from the point of view of sublanguage. It is crucial to determine whether the nursing language can be categorized as a sublanguage. In other words, this research addresses the theory/methodology of determining whether or not some type of language we are looking at is a sublanguage.

This research conducted aims to answer the following research questions:

- Does the language used in Malaysian nursing textbooks and journals have a more restricted lexicon as compared to reference corpora of restricted language and unrestricted language, and if so, in what ways?
- Does the language used in Malaysian nursing textbooks and journals have more restricted morphosyntactic categories as compared to reference corpora of restricted language and unrestricted language, and if so, in what ways?

3. Methodology

Four corpora were collected for this study. The nursing corpus consists of two sub-corpora. The first consists of textbooks used by the nursing students; while the second consists of nursing journals. Another corpus that is included in this study is the weather reports. The weather reports and nursing corpora represent the constrained language in this research. The unconstrained language corpus used for comparison purposes is the BNC Sampler corpus.

Texts in the nursing textbooks are manually extracted and converted into the plain text format. However, certain irrelevant information such as diagrams, tables, references and further reading sections is deleted. In total, the nursing textbook corpus contains 1,005,707 tokens.

Given the importance of nursing research to nursing students and nurse practitioners, nursing journals are also included in this study. The entire issues of three nursing journals ranging from the year 2004 to 2007 were downloaded with the total of 1,051,774 tokens.

In total, 120 weather reports were downloaded for the Metrology Office website. The weather report corpus is then compared to the reference corpus as well as the nursing textbook and nursing journal corpora to measure the degree of closure at the lexical and morphosyntactic. The weather report corpus comprises 193,700 tokens.

For the purpose of comparison, only the written corpus of the BNC (British National Corpus) Sampler will be used as a reference corpus since the other three corpora consist of written data. The BNC Sampler was complied with an aim to maintain comparability with the whole BNC, as well as the integrity of the text samples already in the BNC (BNC Consortium, 2001). For these reasons, the written corpus of the BNC Sampler is chosen as the reference corpus for this study with 1,010,690 tokens.

It is hypothesized that the weather reports, nursing textbooks and nursing journals represent sublanguages. The BNC Sampler was thought to represent unconstrained language.

The current study aims to assess closure at lexical and morphsyntactic levels. To achieve this, five programs or software are used - WordSmith 5.0 (Scott, 2001) and CLAWS (Garside et al., 1987).

4. Findings

4.1. Lexical closure

The examination of closure will begin with lexical closure, which can be examined by determining the type token ratio (henceforth, TTR). From the perspective of TTR, a large TTR signifies a diverse vocabulary (lower degree of closure), while a small ratio suggests a restricted usage of vocabulary

(higher degree of closure). Table 1 shows the type/token ratios of all four corpora at 180,000 tokens.

Table 1: The Type/Token Ratios of all four corpora at 180,000 tokens

Corpus	Tokens	Types	Type/Token Ratio
Weather Report	180,000	3126	0.0172
Nursing Journals	180,000	9467	0.0520
Nursing Textbooks	180,000	9545	0.0521
Bnc Sampler	180,000	15560	0.0863

Based on these ratios, it is evident that the weather report corpus has the smallest TTR; and this is followed by the nursing journals and then nursing textbooks. The BNC Sampler, on the other hand has the largest TTR. This clearly indicates that the weather report corpus is more prone to repetition of words compared to the other three corpora. As for the nursing textbooks and journals, the nursing journals and textbooks show a more restricted vocabulary compared to the BNC Sampler, with the nursing journals more restricted than the nursing textbooks. The evidence suggests that the nursing textbooks have a more diverse lexicon compared to the nursing journals. The BNC Sampler demonstrates the highest TTR suggesting a freer lexical usage with less repetition of lexical forms. The comparison of TTR of all corpora seems to suggest that the BNC Sampler represents unconstrained language while the weather reports represent constrained language.

In lexical terms, the findings seem to suggest that the weather reports have the most restricted lexicon (constrained language) while the BNC Sampler seems to have a freer lexicon (unconstrained language). The nursing journals and textbooks however do not fall in either of these categories. There is a slight difference between the growth rate of the nursing textbooks and nursing journals. Though the decline was small, we should not discard what the findings are pointing to that the lexical growth of the nursing journals is to a slight degree more restricted than that of the nursing textbooks. On the basis of this evidence, this restriction is evidently due to topic restriction. However, the lexicon of these corpora is not nearly as restricted as the weather reports.

Fig. 1 shows running totals of types and tokens of all corpora sampled at the interval of 10,000 tokens. Analyzing the weather reports shows that the curve of the weather report corpus seems to be flattening or to flatten off. The curve represents the lexical growth of the weather reports which indicates that new word forms are encountered least frequently in this corpus. In other words, weather reports display a very restricted lexical usage.

As compared to the nursing textbooks, the nursing journals show a slower lexical growth from the beginning until 1,000,000 tokens. However, there is a slight disparity from the beginning until about 140,000 tokens where the nursing textbooks show a slower lexical growth than the nursing journals.

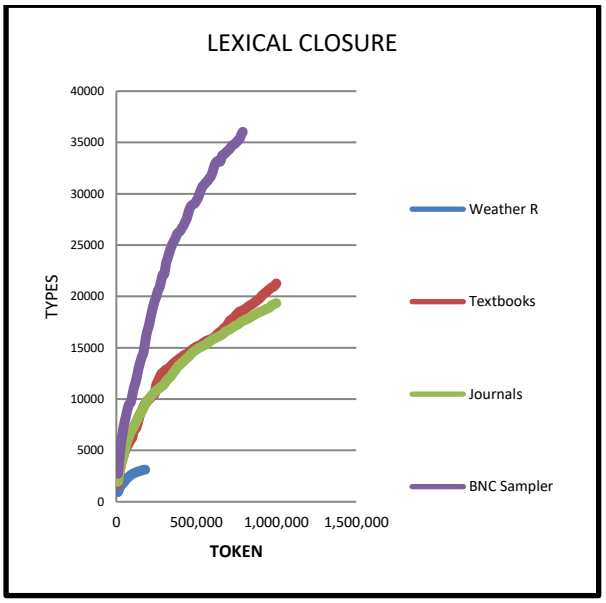

Fig. 1: Lexical growth in all three corpora

From 100,000 tokens to 140,000 tokens the disparity seems greater. Further investigation shows that this disparity is topic-related. Topically speaking, the first 140,000 words of the nursing journals focus on research conducted in the Accident and Emergency Departments. Examples of these types of journals are Accident and Emergency Nursing Journals from the year 2003-2007. The nursing textbooks, however, focus on the basic nursing skills. After the 140,000 token mark, the nursing journals curve moves upward less rapidly and the growth rate reduces slightly indicating that the use of new word forms has become relatively infrequent as compared to the nursing textbooks.

Fig. 1 also shows the curve of the BNC Sampler which seems to display a continuous growth without any sign of approaching closure. This curve denotes a relative high rate of lexical growth in the corpus. This indicates that in terms of lexical growth, new lexical forms in the BNC are encountered most frequently. As the BNC Sampler corpus covers a range of text categories and thus, a great variety of topics, these findings were expected.

Table 2 displays the twenty most frequent lexical words in all the corpora (per million words).

If we compare the lexical words of the four corpora, we can see a significant difference in terms of the frequency of occurrence of lexical words. Looking at the weather reports, we can observe the highest degree of repetition of lexical words in this corpus as compared to the other three corpora indicating the higher degree of closure. For example, the top lexical words in the weather reports are rain (18,006 per million words [henceforth, pmw]), shower (12,206 pmw), north (10,631 pmw). The high frequency of these lexical words clearly shows there are a lot of repetitions in the weather reports with heavy reliance on a small lexicon. The BNC Sampler, on the other hand, has the lowest degree of lexical repetition (lower degree of closure). In the BNC Sampler the top lexical words include only (1,590 pmw), new (1,582 pmw) and said (1,451

pmw). Looking at the frequency of these lexical words, we can see the low of repetition of words. Additionally, Lindquist (2009) explained that the occurrence of lexical words seems to demonstrate the characteristics of each corpus. Therefore, the examination of the top twenty most frequent lexical words in the weather reports reveals that these words are related to weather forecasting in the UK. It is interesting to see that the highest frequent lexis in BNC Sampler, however, is general and does not seem to have a focused subject matter.

Table 2: 20 most frequent lexical words in all three corpora

Weather Reports	Per Mil Word	Nursing Textbooks	Per Mil Word	Nursing Journals	Per Mil Word	Bnc Sampler	Per Mil Word
rain	18006	patient	11172	patients	5617	only	1590
showers	12206	care	2815	care	5145	new	1582
north	10631	blood	2539	study	4905	said	1451
south	9663	use	2518	health	3837	time	1360
west	9375	pressure	2372	nurses	3266	year	1183
east	8775	procedure	2345	data	2895	made	987
dry	7262	tube	2309	research	2790	formula	960
sunny	7219	catheter	2175	participants	2634	people	923
temperatures	6269	skin	2092	nursing	2228	years	918
heavy	6150	equipment	1998	time	2086	system	901
weather	6056	place	1928	used	2035	use	891
recorded	5956	pain	1784	group	2025	number	859
England	5375	sterile	1727	staff	1882	party	847
day	5125	prevent	1722	pain	1591	government	835
pressure	5031	position	1634	use	1590	form	785
average	4906	nursing	1599	reported	1484	do	764
areas	4888	site	1579	women	1481	like	760
sunshine	4569	time	1473	information	1470	back	736
Scotland	4244	remove	1465	clinical	1441	high	665
winds	4175	solution	1408	hospital	1374	make	660

Again, the nursing textbooks and journals fall in between these two categories. In the nursing journals, the top lexical words are patients (5,617 pmw), care (5,145 pmw) and study (4,905 pmw); while the top lexical words in the nursing textbooks are patient (11,172 pmw), care (2,815 pmw) and blood (2,539 pmw). What is interesting to see is the high occurrence of the word patient in the nursing textbooks; which is almost as high as the occurrence of the lexical words of the weather reports. However, judging from the frequency of the second top lexical word care, we can clearly see that the lexical word patient is an outlier signifying the importance of this lexical word to the corpus. Also, what differentiates these two corpora is the occurrence of lexical words such as data, participants and research in the Nursing Journals suggesting that this corpus deals with research process and methods in the nursing domain. Based on these results, we can conclude that there is higher repetition in nursing journals than in the nursing textbooks.

The investigation on lexical closure suggests that the weather report corpus is probably a sublanguage as the corpus does not totally flatten off at the interval of 10,000 tokens. Conversely, the findings also show that the BNC Sampler is more typical of the unrestricted English as the lexical growth curve does not show any sign of flattening off. Words in the BNC Sampler get repeated locally as they are specifically related to the individual sections or articles of the corpus. In the case of the nursing journals and textbooks, neither corpus can be categorized as a sublanguage since none of the lexical growth curves approached closure; but it would be also false to categories these corpora as general English too. There is a high degree of lexical repetition in both corpora. While the lexicons are not as restricted as the weather reports, neither are they as free as the lexis of the BNC Sampler. Therefore, they fall in between these two categories, suggesting the existence of a grey area between the sublanguage and general English. Both the nursing journals and textbooks are related to specialized subject – i.e. the nursing domain. Therefore, these corpora have a specific communicative purpose – conveying domain-specific knowledge to the students/readers. Based on this, it is understandable that the lexicons of these corpora are restricted since they are domain-specific.

4.2. Morphosyntactic closure

The second investigation is directed towards establishing the morphosyntactic behaviour of the words' usage in the corpora studied.

Using the computer POS tagging a finite list of around 50 to 150 separate POS tags can be distinguished (Garside et al., 1987). Due to this, it is inevitable that the growth curves of all four corpora must reach closure because the repertoire of types really is closed (Fig. 2). But plotting the growth curves of all four corpora reveals a similar picture as that in lexical closure. The BNC Sampler has the highest number of different tags as the corpus reached closure at 136 types. The weather reports have the lowest number with 101 types and this is followed closely by the nursing textbooks with 128 types. The nursing journals fall in the third position with 133 types. So on this point, the weather report corpus is the most restricted; while the BNC Sampler uses the greatest number of different syntactic

categories. Table 3 shows the POS tags that cannot be found in the weather reports.

Out of 137 POS tags in the CLAWS 7 TAGSET, 36 tags are not found in the weather reports. For examples in terms of pronouns, in the weather reports pronoun tags including the first person singular objective personal pronoun (me), first person plural objective personal pronoun (us), first person singular subjective personal pronoun (I) and first person plural subjective personal pronoun (we) are absent. Verb tags that are absent in the weather reports are the base form of be (finite i.e. imperative, subjunctive), am, base form of do (finite) and doing. This is clear evidence that the weather reports, as expected for a sublanguage, are not using the full grammatical resources of the language. Next, a comparison of the Nursing Journals and Textbooks in Table 4 indicates interesting similarities and differences.

Tags that cannot be found in either the nursing journals or the textbooks are MCGE, NNL2, NPM2, VVNK and ZZ2. By contrast, tags that can only be found in the nursing journals are NNA, NNL1 and UH. Table 5 shows some examples of these tags found in the nursing journals.

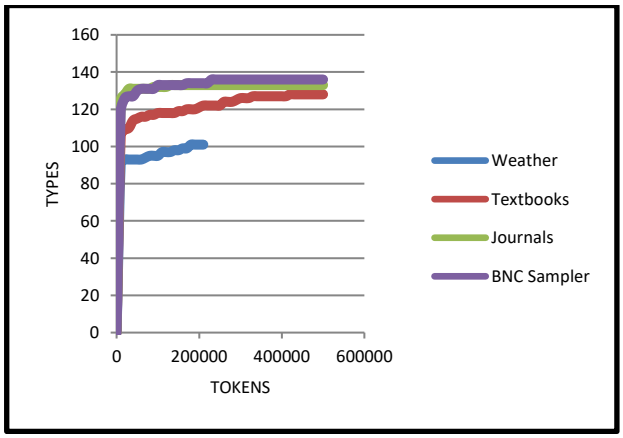

Fig. 2: POS Growth in all corpora

Table 3: POS that are not found in the weather reports

CLAWS 7 TAGSET	
BCL	before-clause marker (e.g. in order (that),in order (to))
CSW	whether (as conjunction)
DDQV	wh-ever determiner, (whichever, whatever)
FO	formula
FW	foreign word
MCGE	genitive cardinal number, neutral for number (two's, 100's)
MCMC	hyphenated number (40-50, 1770-1827)
NNA	following noun of title (e.g. M.A.)
NNB	preceding noun of title (e.g. Mr., Prof.)
NP	proper noun, neutral for number (e.g. IBM, Andes)
NPD2	plural weekday noun (e.g. Sundays)
PNQO	objective wh-pronoun (whom)
PNQS	subjective wh-pronoun (who)
PNQV	wh-ever pronoun (whoever)
PNXI	reflexive indefinite pronoun (oneself)
PPGE	nominal possessive personal pronoun (e.g. mine, yours)
PPHO1	3rd person sing. objective personal pronoun (him, her)
PPHS1	3rd person sing. subjective personal pronoun (he, she)
PPIO1	1st person sing. objective personal pronoun (me)
PPIO2	1st person plural objective personal pronoun (us)
PPIS1	1st person sing. subjective personal pronoun (I)
PPIS2	1st person plural subjective personal pronoun (we)
PPY	2nd person personal pronoun (you)
RPK	prep. adv., catenative (about in, be about to)
UH	interjection (e.g. oh, yes, um)
VBO	be, base form (finite i.e. imperative, subjunctive)
VBM	am
VDO	do, base form (finite)
VDG	doing
VDI	do, infinitive (I may do... To do...)
VDN	done
VDZ	does
VVGK	-ing participle catenative (going in be going to)
VVNK	past participle catenative (e.g. bound in be bound to)
ZZ2	plural letter of the alphabet (e.g. A's, b's)

Table 4: Comparison of POS between the nursing journals and textbooks

	CLAWS 7 TAGSET	Journals	Textbooks
MCGE	genitive cardinal number, neutral for number (two's, 100's)	NO	NO
NNA	following noun of title (e.g. M.A.)	YES	NO
NNL1	singular locative noun (e.g. Island, Street)	YES	NO
NNL2	plural locative noun (e.g. Islands, Streets)	NO	NO
NPM2	plural month noun (e.g. Octobers)	NO	NO
UH	interjection (e.g. oh, yes, um)	YES	NO
VVNK	past participle catenative (e.g. bound in be bound to)	NO	NO
ZZ2	plural letter of the alphabet (e.g. A's, b's)	NO	NO

Table 5: Examples in the nursing journals

TAG	NURSING JOURNALS
NNA	at_II William_NP1 R._NP1 Sharpe_NP1 ,_, Jr_NNA ._. Hospital_NN1
NNL1	the_AT Hampshire_NP1 and_CC Isle_NNL1
UH	Oh_UH well_RR ,_, they're_NN1 intubed_VVD ._.

Academic journals contain scholarly works that introduce and present new research or critique existing research. Therefore, it is surprising to find the tag UH (interjection - e.g. oh, yes, um) in the nursing journals. Further investigation of the nursing journal corpus revealed that the occurrence of this tag (UH) is due to the presentation of qualitative findings of various researches, particularly research which involves interviewing the participants. For example, the phrase 'Oh well, they're intubed.' is from an article entitled 'Communication boards in critical care: patients' views'. This research involves identifying the level of frustration of patients who received mechanical ventilation. To achieve this 29 patients were interviewed. The use of interjections occurred when the results of the interviews were presented in these articles.

On the other hand, it is not surprising that this list of tags (NNA, NNL1 and UH) cannot be found in the nursing textbooks. The nursing textbooks serve a variety of purposes. One of the prime purposes of these textbooks is becoming a manual of instruction in the nursing domain. Thus, it would be considered strange to find interjections (UH tag) in nursing textbooks. The same goes for the tag NNL2 (plural locative noun - e.g. Islands, Streets)

The findings prove that the POS types of the weather reports are the most limited; while the BNC Sampler contains the most diversified POS types. The shapes of the growth curves of all four corpora are not especially interesting since the curve flattens off quickly due to the closed repertoire of POS types. However, a comparison of the POS types in the four corpora has produced an interesting picture. It shows that the nature of a corpus dictates the presence or absence of certain POS tags, which is an indicator of the extent to which all the grammatical resources of a language have been exploited.

The next investigation explores the POS tags attached to each word in the four corpora at the 180,000 token intervals. Though this analysis may seem similar to that of the lexical closure, it is crucial to understand that the Word/POS analysis is an extension to the lexical closure as the analysis of the Word/POS involves looking at the POS tags attached to each word. We know in advance that a big part of the POS/Word growth is actually the lexical growth. However, it would be of considerable interest to know the extent of this growth as compared to the lexical growth. The analysis of Word/POS will enable us to determine the grammatical flexibility or grammatical rigidity of words within each corpus. A large TTR, therefore, shows that the grammatical flexibility of the corpus is high as the words in the lexicon of the corpus are associated with a wider variety of syntactic functions. A small ratio, however, indicates grammatical rigidity of the corpus as the syntactic functions that are associated to each word are fewer.

At 180,000 tokens the BNC Sampler corpus has a TTR of 0.0761 while the weather reports have the lowest ratio of 0.0171 (Table 6). The nursing textbooks and journals have similar ratios of 0.0495 and 0.0504, respectively. The evidence (in comparison the figures for lexical closure) clearly shows that the words in the BNC Sampler have a wider variety of word-tag pairings (high TTR), but just the reverse is true for the weather reports (low TTR). Rather, the word-tag pairings of the weather report corpus seem decidedly finite. In other words, the lexicon in the BNC Sampler shows greater grammatical flexibility than the weather reports. Conversely, we can see a small difference between the ratios of the nursing textbooks and journals and this is in contrast to the findings obtained for the lexical closure (see Section 4.1). Here, the lexicon in the nursing journals is associated with a wider variety of word-function pairings than the nursing textbook.

Table 6: Word/POS Tag ratio in all three corpora at 180,000 tokens

CORPUS	SIZE	TYPES	TYPE/TOKEN RATIO
WEATHER REPORTS	180,000	3094	0.0171
NURSING TEXTBOOKS	180,000	8915	0.0495
NURSING JOURNALS	180,000	9069	0.0504
BNC SAMPLER	180,000	13700	0.0761

Moving ahead, growth curves are then plotted to determine closure at the level of word-tag pairings (Fig. 3). This is essential as a low TTR does not necessary mean that morphosyntactic closure has occurred. Additionally, plotting the growth graphs will enable us to determine how often the lexicon in each corpus is associated with different POS. Therefore, closure at the morphosyntactic level becomes evident if the growth curve of new type+tag combinations flattens off; while the reverse will happen for unrestricted language as the growth curve will continue without flattening towards to horizontal. Since only the weather report corpus is predicted to reach closure at the lexical level, it is necessarily true that the similar findings will be obtained at the morphosyntactic level as there cannot be fewer type+tag pairs than word types. Hence, the prime aim of this analysis is not to see whether any of the corpora under study reaches closure at morphosyntactic level, rather it is to look for the differences or similarities between the

word/POS curves and the lexical curves in the previous section (Fig. 1).

Fig. 3 illustrates the word/POS growth curves of the corpora. The findings clearly show that none of the growth curves level off. In other words, none of the corpora reaches closure at morphosyntactic level, as expected. All the main corpora continue to find different syntactic uses for already-used lexis quite steadily. This is not true of the weather report corpus. But, with that exception, the other corpora are adding new lexicon and more importantly, as these words are added, new syntactic functions of these words are also added. It appears that the lexis in the weather reports are association with limited POS indicating the restricted grammatical flexibility of the corpus.

A comparison of the growth curves of the nursing textbooks and journals reveals an interesting story. In these corpora, there seems to be a similarity in how often new POS tags are associated with already-seen types. Though neither of the nursing textbooks and journals growth curves flattens off, these curves appear almost identical. Rate at which new word+tag combinations are observed in the nursing journals and textbooks seems equal. What is important is the fact that, just like the findings in the lexical closure, the growth curves of both nursing journals and textbooks are more restricted than the BNC Sampler. The growth curve of the BNC Sampler continues to develop showing no sign of levelling off. This indicates that the lexicon of the BNC Sampler can be associated with a variety of syntactic uses.

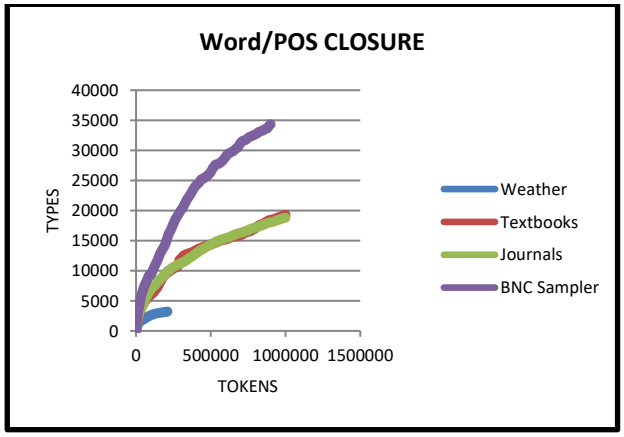

Fig. 3: Word/POS growth in all corpora

The next step is to compare the lexical and the Word/POS growth of the four main corpora (Figs. 4, 5, 6 and 7). Fig. 4 shows the lexical and POS growth curves of the weather reports. Looking at the gap between these lines, we can see that the difference between them is very small. This shows that the grammatical flexibility of the words in the corpus is very limited. Words are typically not reused in novel grammatical categories.

In contrast to the findings obtained in the weather reports, the gap between the lexical and the Word/POS lines in the BNC Sampler is greater (Fig. 5). Comparing the weather reports and the BNC Sampler, we can clearly see the grammatical flexibility of words in the BNC Sampler.

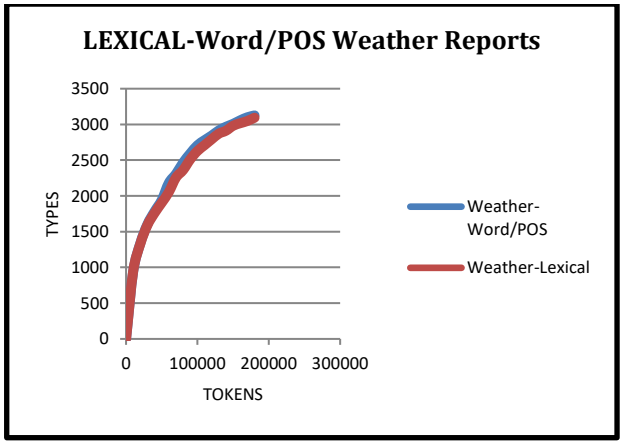

Fig. 4: Comparison between lexical and Word/ POS in weather reports

As for the lexical and Word/POS growths of the nursing textbooks and journals, we can see the difference between them (Figs. 6 and 7). The difference between the lexical and Word/POS curves of the nursing textbooks is fairly large indicating that the words in the corpus have multiple grammatical categories. The difference between the lexical and Word/POS growths in the nursing journals, on the other hand, seems quite small reflecting a restricted grammatical flexibility of the corpus. What is interesting to see by comparing the lexical and Word/POS curves of the nursing textbooks and journals, the nursing textbooks seem to have more grammatical flexibility than the nursing journals.

The first 200,000 tokens of all four main corpora were analyzed to determine the rate of lexical growth for each main category of lexical words; namely Nouns, Verbs, Adjectives and Adverbs (Figs. 8, 9, 10 and 11).

Looking at the growth curves, two obvious points emerge. First, it is clear that all the growth curves seem to paint the same picture – i.e. the weather reports are the most restricted with the least number of types; while the BNC Sampler has the most types. The nursing textbooks and journals again fall in the middle and close together. Secondly, based on these graphs the closing off of the weather reports becomes much more obvious.

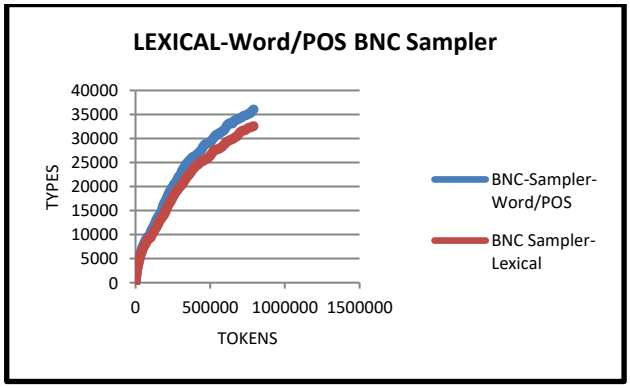

Fig. 5: Comparison between lexical and Word/POS in the BNC sampler.

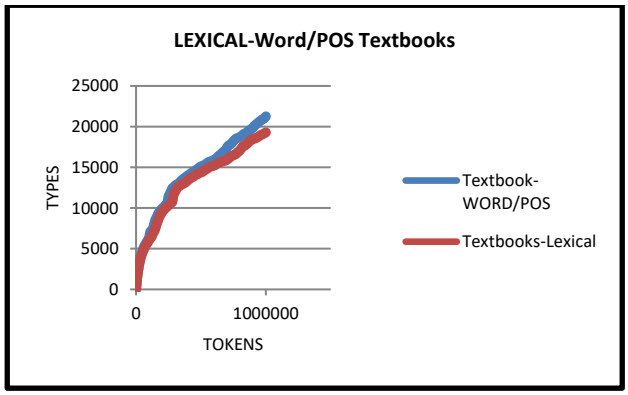

Fig. 6: Comparison between lexical and Word/POS in nursing textbooks

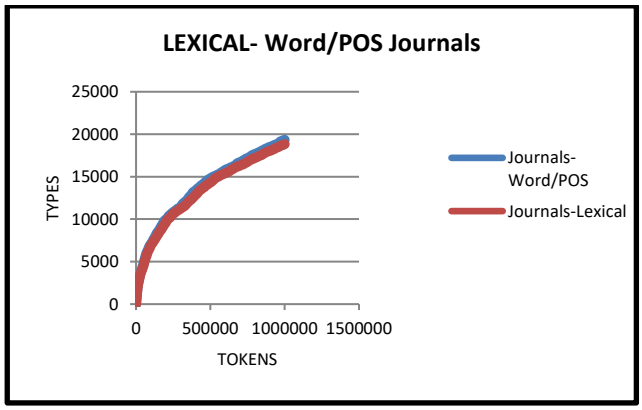

Fig. 7: Comparison between lexical and Word/POS in nursing journals

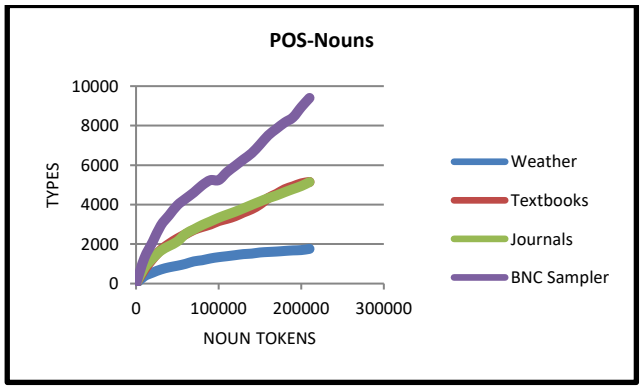

Fig. 8: Nouns in all corpora

However, looking at the first three graphs we can see notable differences (Figs. 8, 9 and 10). The graphs show that the Noun types in the nursing textbooks and journals are much more closed than the BNC Sampler corpus. Yet, when we compare the growth curves of the Verb and Adjectives types in these corpora a different finding emerge – in this respect they seem to be almost as free as the BNC Sampler.

5. Conclusion

The foregoing analyses showed that the only corpus that reached closure at the lexical and morphosyntactic levels was the weather reports. As for the nursing textbooks and journals, the investigations showed that they did not approach closure at any level. The BNC Sampler was behaving exactly as unconstrained language is expected to, not approaching closure at any level. So, the findings of this study did not point to an affirmative answer to the main research question. It had been hypothesized that the datasets other than the BNC Sampler represent sublanguages. The results of the investigations, however, do not support this hypothesis. Only the weather reports corpus can be categorized as sublanguages. Further analyses at n-grams and constituent levels should be conducted to deduce that all corpora can be unambiguously categorized as sublanguages.

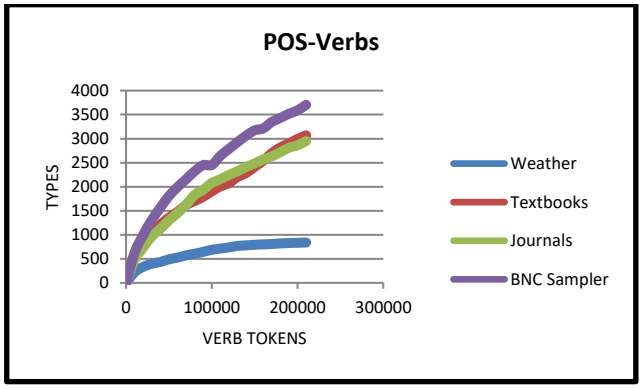

Fig. 9: Verbs types in all corpora

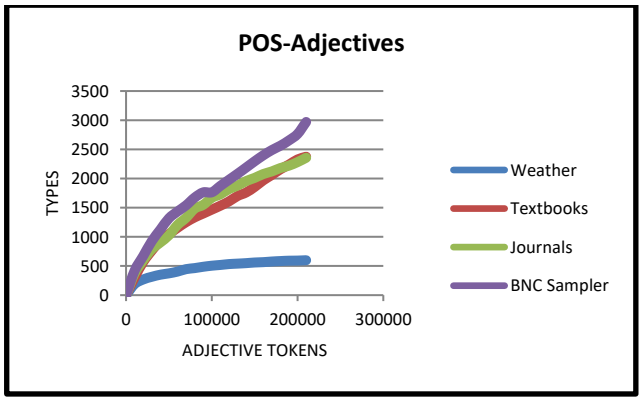

Fig. 10: Adjectives in all corpora

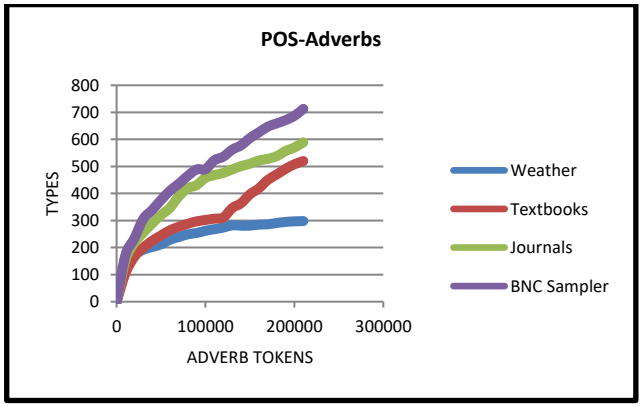

Fig. 11: Adverbs in all corpora

This study was conducted with the chief purpose of investigating closure in the nursing textbooks and the nursing journals. Interestingly, the findings of the study reveal that the idea of a sublanguage is problematic.

The findings of the study seem to cast a doubt on the definition of a sublanguage. Based on the

definitions of sublanguages, it was hypothesized that the nursing textbooks and journals can be categorized as a sublanguage as they are highly specialized and utilized by trained healthcare personnel in the domain of nursing. This is the sort of language that we would expect to be a sublanguage, according to the criteria in the language. But the investigations carried out on the nursing textbooks and journals do not seem to yield the expected results. Though these corpora have domain specific lexis and morphosyntactic, none of the growth curves of these corpora approach closure. The findings, in fact show that the corpora seem to belong in a middle area between highly constrained language and highly unconstrained language. So clearly, not all highly specialized domains can be automatically categorized as sublanguages.

The findings also show that that closure is not a unitary phenomenon. As demonstrated in this research, none of the growth curves of the nursing textbooks and journals approach closure at any level. Based on this evidence we can conclude that the presence of some degree of restriction in the corpora at the lexical or morphosyntactic levels does not ensure complete closure. The original definition of a sublanguage seems to clearly divide a sublanguage or constrained language from unconstrained language and placing both as two discrete dichotomies. However, the findings regarding the nursing textbooks and journals seem to show that there is no explicit or clear-cut boundary that divides constrained language from unconstrained language. Instead, it shows that the concept of sublanguage should be on a continuum, with constrained language/sublanguage and unconstrained languages at the two extremes ends. Rather than looking at these extremes as discrete categories, another area should be added to represent domains such as the nursing textbooks and nursing journals that are not as restricted as constrained language, but not as free/flexible as unconstrained language. This newly added area should be categorized as partly constrained language (Fig. 12). This area, however, is fuzzy and relative on the cline. Based on the evidence, the nursing textbooks and nursing journals should be placed in the middle of this continuum to represent partly constrained language, since neither of these corpora could be categorized as a sublanguage, but they are clearly more restricted in many respects than wholly unconstrained language as found in the BNC Sampler.

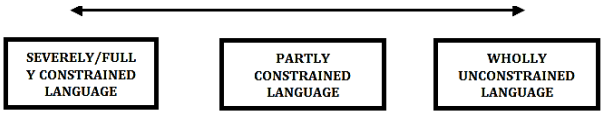

Fig. 12: The idea of sublanguage

The findings also show that there are various factors that may position a type of language between these two extremes, including domain specific lexis, highly specialized formulae, and stereotypical constituent structures. The evidence shows that these factors determine whether one domain is more restricted than another. The weather report corpus is still the most restricted corpus in this study since there are high incidences of domain specific lexis and morphosysntax. Thus, the weather report corpus should be intermediate between the fully constrained and partly constrained language with respect to the reasons outlined above. The nursing textbooks are a type of partly constrained language, less restricted than the weather reports in many respects but it is not as free/flexible as unconstrained language. Fig. 13 outlines the continuous parameters of variation between the constrained and unconstrained language; this further strengthens our stance that constrained and unconstrained language should not be seen as a simple dichotomy. Instead it should be in a continuum governed by the parameters noted above.

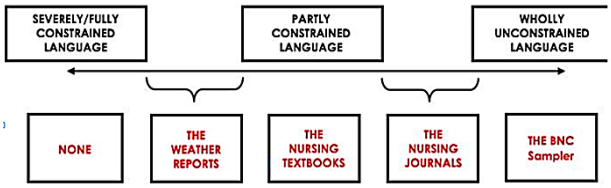

Fig. 13: The positions of each corpus in a continuum

References

BNC Consortium (2001). British National Corpus (BNC world Edition, Version 2). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Available online at: http://www.natcorp.ox.ac.uk/

Bross IDJ (1972). How information is carried in scientific sub-languages. Science, 176(4041): 1303-1307.

Deville G (2001). Corpus-based sublanguage modelling for NLP applications: a tentative methodology. In the International Colloquium on Trends in Special Language and Language Technology, Brussels, Belgium.

Erickson F (1986). Qualitative Methods in Research on Teaching. In: Wittrock MC (Ed.), Handbook of Research Teaching: 119-161. 3rd Edition, Macmillan, New York, USA.

Friedman C, Kra P, and Rzhetsky A (2002). Two biomedical sublanguages: a description based on the theories of Zellig Harris. Journal of Biomedical Informatics, 35(4): 222-235.

Garside R, Sampson G, and Leech G (1987). The computational analysis of English: A corpus-based approach. Longman, London, UK.

Harris Z (1968). Mathematical structures of language. Wiley, New York, USA.

Hutchinson T and Waters A (1987). English for specific purposes. Cambridge University Press, Cambridge, UK.

Lindquist H (2009). Corpus Linguistics and the Description of English. Edinburgh University Press, Edinburgh, Scotland.

McDonough J (1984). ESP in perspective: A practical guide. Collins ELT, London, UK.

McEnery T and Wilson A (1996). Corpus Linguistics. Edinburg University Press, Edinburg, USA.

Sager N (1986). Sublanguage: Linguistic phenomenon, computational tool. In: Grishman R and Kittredge R (Eds.), Analyzing language in restricted domains: sublanguage description and processing: 1-16. Lawrence Erlbaum Associates Publishers, London, UK.

Scott M (2001). Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith Tools suite of computer programs. In: Henry MA and Roseberry LR (Eds.), Small Corpus Studies and ELT: 47-67. John Benjamins, Amsterdam, Netherlands.

Spyns P (1996). Natural language processing. Methods of Information in Medicine, 35(4): 285-301.

Travers DA and Haas SW (2003). Using nurses' natural language entries to build a concept-oriented terminology for patients' chief complaints in the emergency department. Journal of Biomedical Informatics, 36(4): 260-270.